

Attachment A**Application Cover Sheet**

Cleartelligence Holdings, LLC			
Mailing Address 275 Grove Street Suite, 2-400	City/Town Newton	State MA	Zip Code 02466
Telephone 617-340-7740	Fax 617-340-7740	Web Address cleartelligence.com	
Primary Contact for Clarification Kevin Lazorik		Primary Contact E-mail Address kevin.lazorik@cleartelligence.com	
Authorized Signatory Kevin Lazorik		Authorized Signatory E-mail Address kevin.lazorik@cleartelligence.com	
Legal Status/Jurisdiction – a Massachusetts Corporation, LLC		Respondents W-9 (Attached) Respondents UEI No. J8BLJ6MYDSY5	

Proposal Contents

Letter from Cleartelligence.....	6
Executive Summary	7
Our vision for the Data Commons Collaborative	7
Perfect fit for your success	7
Timeline & budget (at-a-glance).....	8
A partner for your success	8
Organizational Overview & Qualifications	9
Cleartelligence Overview	9
Why Cleartelligence for the Data Commons Collaborative	9
Core Capabilities Alignment	9
Public Sector Experience	10
Case Studies & References.....	10
Approach & Methodology.....	12
Program Delivery.....	12
Program Management	12
Team Structure.....	13
Audience Engagement.....	14
Technical Approach & Architecture	16
Web Experience & Data Management – PortalUS + CKAN	16
Security & Authentication	20
Metadata Standards and Interoperability	25
Common Credits.....	26
Warehousing & Hosting	30
Logging & Centralized Observability	35
Reporting & Dashboards.....	39
Dataset Curation, Integration, & Governance	40
Content Management.....	44
Translation & Internationalization	46
AI Fairness and Bias Framework	50
Synthetic Data Capability	53

Phases, Scoped Activities, and Deliverables.....	55
Phase 1 – Initiation & Design.....	55
Strategy and Requirement Alignment	55
Detailed Requirement Gathering.....	56
Program Planning	56
Phase 2 – Infrastructure & Setup.....	57
Environment and 3 rd Party Service Setup	57
Experience Design	57
Phase 3 – Development.....	58
Agile Solution Development	58
Phase 4 – Testing & Governance	59
QA & Performance Testing	59
Security & Penetration Testing	60
User Acceptance Testing (UAT) with Stakeholders & Pilot Users	60
Phase 5 – Rollout & Adoption.....	61
Product Launch.....	61
Post Launch Monitoring & Warranty	61
Recommended Services & Platforms	63
Solution Analysis & Justification	63
Comprehensive Knowledge Archive Network (CKAN)	63
Data Management Platform Comparison	65
Auth0.....	65
Okta vs. Auth0 by Okta	67
bridge.xyz.....	67
Metabase.....	69
Snowflake	70
Role of Snowflake in the DCC Architecture.....	70
Strapi.io	71
Benefits of using Strapi.io	71
Google Analytics	72
Translations.com	73
Alignment to DCC Requirements	73

Team & Key Personnel	75
Chitra Sundaram - Practice Director, Data Management.....	75
John Fico – Practice Director, Software Engineering	76
Pete Peralta – Managing Director	77
Keith Meyer – Practice Leader, Artificial Intelligence	78
Lisa Iyer – Principal Architect	79
Brian Bailey – Senior Consultant, Artificial Intelligence	80
Mason Reiter – Consultant, Artificial Intelligence	81
Timeline and Budget	82
Engagement Timeline	82
Budget Summary	83
Services Budget.....	83
Software, Hosting, & SaaS Budget.....	85
Additional Options to Manage Budget.....	87
Offshore QA	87
Offshore QA + Limited Offshore Development.....	87
Case Studies	88
Teradyne.....	88
Kiniksa	89
Safe Schools for Alex.....	90
Compliance and Acceptance.....	91
Clearintelligence W9	92
Appendix	93
Index of Figures	93
Links to Individual Files	94
Key Assumptions & Dependencies	95
Solutions & Licensing	95
Experience & Branding.....	95
Program Governance & Oversight	95
Performance & Testing	95
Industry Engagement Plan Framework.....	96
Goals	96

Engagement Activities	96
Initial KPIs for Discussion.....	97
Sample Fairness Report	98
Overview	98
Dataset Under Review	98
Methodology	98
Findings	98
Fairness Impact Statement.....	99
Dashboard Screenshot (Illustrative).....	99
Supporting Technical Information.....	100
Metadata Field Mapping	100
Commons Credits Ledger Schema Example	101
CMS + CKAN Content Model.....	102

Letter from Cleartelligence

October 7, 2025

Dear Sabrina and Members of the RFP Review Team,

On behalf of Cleartelligence, I would like to express our genuine interest in this project and our appreciation for the opportunity to submit a proposal in response to *RFP No. 2026-JAII-01: AI Hub Data Commons Consulting Services*. Thank you for the time and effort you and your colleagues are dedicating to reviewing proposals for this important initiative.

Cleartelligence has spent more than 15 years helping organizations in both the public and private sectors harness the power of data and AI to drive innovation, improve decision-making, and deliver measurable impact. Our extensive experience across regulated industries, government agencies, and mission-driven institutions uniquely positions us to support the Massachusetts Technology Collaborative in designing and delivering the Data Commons Collaborative. As a Massachusetts based business, we know the capabilities the Commonwealth offers through its unique combination of government investment, leading higher education institutions, and technology innovation culture.

By choosing our firm, your team gains a partner that shares its mission to accelerate innovation, strengthen equity, and build public trust in AI. Like you, we serve the explorers, innovators, and builders who turn ideas into impact by providing them with the data, tools, and expertise needed to create responsible AI solutions.

Our experience in designing secure, scalable platforms, enabling synthetic data for experimentation, and embedding fairness into AI development aligns directly with the Data Commons' vision. Just as the Commons will fuel collaboration and equitable access across Massachusetts, our work empowers diverse stakeholders to discover insights, test solutions, and bring transformative technologies to market.

Together, we can translate MassTech's strategy into a practical, future-ready infrastructure that not only meets today's needs but also positions Massachusetts as a global leader in ethical and impactful AI innovation.

Thank you for considering our proposal and for your leadership in advancing AI innovation in Massachusetts. We look forward to the possibility of working together on this transformative initiative.

Sincerely,



Owen Frivold
Chief Executive Officer

cleartelligence 

Executive Summary

Our vision for the Data Commons Collaborative

We see the DCC as Massachusetts' front door to responsible AI: a secure, extensible hub that makes high-quality data easy to find, safe to use, and energizing to build with. The platform centralizes access to curated public and Massachusetts datasets, integrates synthetic data generation for privacy-preserving experimentation, and bakes in fairness, governance, and auditability from day one. Researchers, startups, agencies, and industry teams get a single, welcoming place to discover datasets, preview them, request access, generate synthetic analogs, and track usage via a transparent Commons Credits ledger, accelerating real-world AI adoption while protecting people, privacy, and policy.

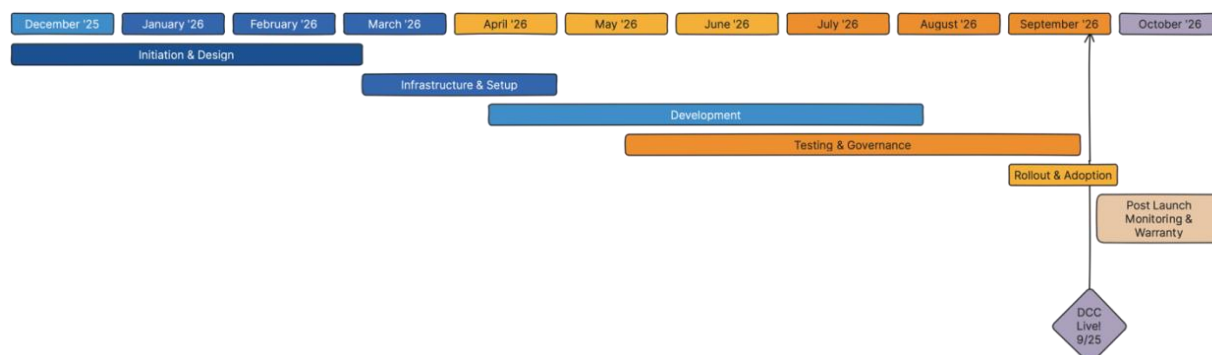
Perfect fit for your success

We're a Massachusetts headquartered company that's spent 15+ years shipping data and AI solutions for mission-driven organizations. This program plays to all our strengths:

- **Public-sector muscle memory.** We know governance, privacy, and accessibility expectations—and how to meet them without slowing teams down.
- **Design + build + adopt.** We don't just stand up tech; we curate initial datasets, define clear governance, and enable administrators and contributors so the DCC grows sustainably.
- **Proven stack, zero drama delivery.** CKAN for catalog and metadata standards (DCAT/Dublin Core/Schema.org), Auth0 for Security, and an analytics layer for usage, audit, and reporting paired with synthetic tooling (e.g., SDV/Synthea/CARLA) and fairness tooling (AIF360/Fairlearn) so equity and safety aren't afterthoughts.
- **Flexible and focused.** Our expert practitioner-led teams are lean, agile, and adept at delivering the results clients need with both speed and high-touch service. This outcomes-first model has given us a front-row seat to solving complex, data-forward problems for some of the largest public and private organizations in the world.
- **Certified AI Nerds.** We're continuously training and upskilling our team through knowledge sharing and formal AI training, evolving how AI integrates into our service delivery, and creating internal AI sandboxes for practitioners to safely experiment with tools and solutions. Our commitment to building an AI-ready culture, deep technical expertise, and a continuous learning environment equips us to stay on top of fast-moving AI tools and trends, so you don't have to.

Timeline & budget (at-a-glance)

We propose a phased delivery with regular show-and-tell checkpoints (every ~3 weeks) from kickoff through public launch and post-launch support. With a kickoff on 12/2, the DCC will launch in September 2026.



The budget is structured by phase so the Massachusetts Technology Collaborative (MassTech) can manage pace and emphasis with full transparency. Delivery will be done by our US based team, but options for a blended team are provided and could lower the total services budget by over \$1M if desired.

Phase	Hours	Fee
Initiation & Design	6,220	\$1,461,700
Infrastructure & Setup	4,960	\$1,165,600
Development	10,197	\$2,396,295
Testing & Governance	2,639	\$620,165
Rollout & Adoption	1,650	\$387,750
Warranty	360	\$84,600
Total	26,026	\$6,116,110

Software and hosting costs are estimated to start at \$43,000 per year.

A partner for your success

You're not just buying hours; you're gaining a team that will co-own outcomes. We'll keep the work visible, decisions documented, and stakeholders engaged. We'll align governance with real usage (not paperwork theater), deliver quick wins while laying a strong foundation for growth, and measure what matters: adoption, dataset quality, safe reuse, and time-to-impact for AI projects. We know this program will evolve and mature over time, so we design for sustainability, ensuring today's wins compound into tomorrow's capabilities. When questions get hard—privacy, fairness, sustainability—we bring pragmatism and proof points.

The DCC will make Massachusetts the easiest place in the country to do responsible AI. We're excited to build it with you—and to make sure it thrives long after launch.

Organizational Overview & Qualifications

Cleartelligence Overview

Cleartelligence is a Massachusetts-based data and AI consulting firm with deep expertise helping public sector, higher education, and regulated organizations advance their data infrastructure, implement governance frameworks, and leverage artificial intelligence to deliver public value. We combine cutting-edge technical skills with domain knowledge of government operations, compliance, and equity considerations. In July 2024, we acquired the Bardess Group, a complementary group of talented data innovators with deep federal government experience¹.

Why Cleartelligence for the Data Commons Collaborative

The creation and launch of the DCC is a large program with numerous stakeholders, integrations, and customizations that introduce complexity and risk. This is our favorite type of engagement!

- We turn your data chaos into clear and actionable strategies, architectures, and management approaches that ensure data is secure, accessible, and ready to drive business growth.
- From strategy to implementation and beyond, we integrated business, data, and artificial intelligence solutions that users across your organization can trust to make smarter decisions faster.
- Our experienced consultants use their deep industry and technical expertise to deliver scalable, secure and innovative data and AI solutions that are focused on impact, not hype
- As a practitioner-led consultancy, we deliver with agility, depth, and purpose that larger firms cannot match. Our senior experts will build, not just advise. We move fast, stay close to stakeholders, and create practical, trustworthy AI solutions that serve Massachusetts and the public good.

Cleartelligence brings the technical depth, public sector experience, and commitment to ethical AI necessary to support the Massachusetts AI Hub in launching a world-class Data Commons Collaborative.

Core Capabilities Alignment

Our core services align to the broad skillsets needed to fully deliver the Data Commons Collection.

- **Experience Design** – Data Portals are as effective as their design allows. We turn complex requirements into clear, intuitive designs that guide users to action with confidence. By combining information architecture, wireframing, and visual design best practices, we create experiences that are not only engaging but also easy to understand and act upon.

¹ As we integrated Bardess into the core organization, both entities NOW utilize the same TIN and DUNS identifier. Our UEI number is still listed under “Bardess Group, Ltd” and we are in the process of migrating the registration to the Cleartelligence name.

- **Data Strategy, Governance, and Advisory** – We turn data chaos into clarity with roadmaps, governance frameworks, and support that address technology, processes, and people.
- **Data Engineering, Security and Architecture** – We don't just engineer data, we design for long-term impact to ensure your data is secure, accessible, and ready to drive results with the robust data architectures and management strategies we design and activate.
- **AI & Analytics Enablement** – We don't just build bots, we decode your organization first to deploy AI/ML models, synthetic data generation tools, and advanced analytics to improve decision-making.
- **Change Management & Sustainability** – Provide training, documentation, and governance templates to ensure long-term adoption and value.

Public Sector Experience

We service stakeholders across multiple verticals and organizations. This experience brings us engagement with various leaders who align to both the goals and the DCC and the diverse audience of stakeholders you wish to engage:

- **Government & Federal Agencies** - Completed work with the U.S. Department of Defense (details confidential), delivering secure data integration and AI enablement under strict compliance requirements.
- **Non-Profit** – Created and continue to expand the most comprehensive database of school safety incidents in America's schools through our partnership with [Safe Schools for Alex](#).
- **Higher Education** – Extensive relationship with Harvard Business Publishing, a non-profit subsidiary of Harvard University, started in 2021 and spans from enablement and data governance for Harvard Business Review to Training & Enablement to predictive forecasting and project profitability.

Case Studies & References

Each use case links to a long form version in the [Case Studies](#) section of this document.

Teradyne – Manufacturing Planning & Slot Simulation Portal

[Teradyne](#) is a global supplier of automated test equipment (ATE) and industrial automation solutions, based in North Reading, MA. We developed a portal that allows users to upload specifications for what they need to build for a customer. For each submission, the user can log in to a simulation experience to understand what time options would be for the build and subsequent delivery. If they don't find the timeline they are looking for, they can swap others out of a manufacturing slot and see what the impact would be. Impacts considered are not only work center time but also bill of materials and inventory/parts lead time. Once the implications of a change are understood, a committee gathers to review impacts and accept or reject individual changes.

Kiniksa – Clinical Trials Data Management System

[Kiniksa Pharmaceuticals](#) is a biopharmaceutical company focused on discovering, developing, and commercializing therapies for patients with autoinflammatory and autoimmune diseases. They are based in Lexington, MA. We developed a configurable file intake and download system that enables the company to set-up partner companies (e.g. Contract Research Organizations) to upload clinical trials data. Partners are assigned API credentials and pipeline codes for different inputs. When invoked by a partner, the system takes in the file, stores it, logs it, runs it through validation, and invokes a configured pipeline to a target database destination. The partner company gets a success or error response. The API's allow for browsing history of file submissions. The system has internal management features like notification when an expected file is not received.

Safe Schools for Alex – School Safety Data

[Safe Schools for Alex](#) is a nonprofit organization founded in memory of Alex Schachter, one of the 17 students and staff killed in the 2018 Marjory Stoneman Douglas High School shooting in Parkland, Florida. The organization is dedicated to improving school safety nationwide through data-driven research, advocacy, and resources for parents, educators, and policymakers. We have the honor of [supporting them in their mission](#) to ensure every child has a safe and supportive learning environment through their [School Safety Dashboard](#) initiative.

References

As requested, below are three reference contacts who can speak to the talent on our team, the diligence we apply to our work, and the quality focus we bring to all deliverables. Out of respect for each individual's time, we ask that you contact these individuals if we are being considered for advancement into the next round of selection.

Safe Schools for Alex

Max Schachter, Founder & Executive Director
Max@AlexSchachter.org

Boston Scientific Corp

Kaitlyn Donovan, Senior Director, Data Enablement
Kaitlyn.Donovan@bsci.com

Harvard Business Publishing

Maria Herrera, Director, Data Engineering
mherrera@harvardbusiness.org

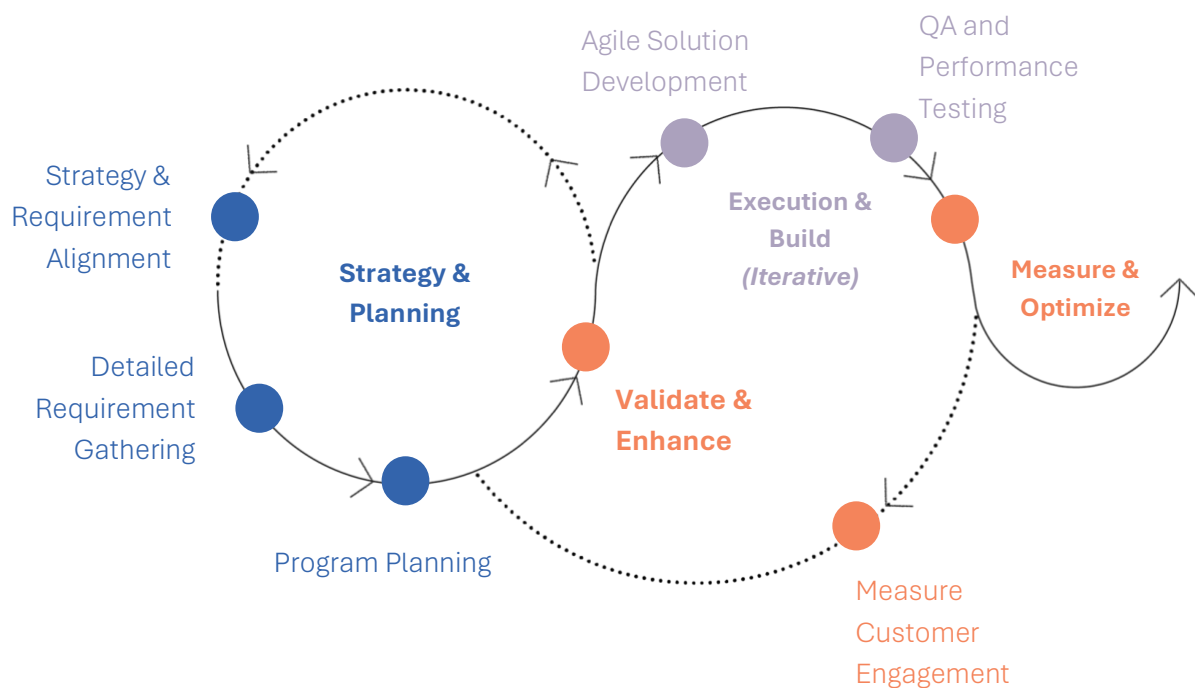
Approach & Methodology

Program Delivery

Our modified approach fully aligns with the five program phases outlined in the Massachusetts AI Hub Data Commons RFP, while also augmenting them with added value. We recommend early user research and piloting to validate direction, agile and iterative planning to adapt to evolving needs, and embedded fairness and governance to ensure trust.

By layering in these enhancements, we aim to deliver not only a compliant platform but also a sustainable, user-centered Data Commons that drives long-term adoption and measurable impact across the Commonwealth's AI ecosystem.

Our hybrid methodology (waterfall + agile) establishes a foundation for an ongoing product lifecycle where measurement drives evolution. User adoption, dataset usage, and fairness audits generate insights that feed directly into the backlog, which is then prioritized and delivered through iterative development cycles. This ensures the Data Commons continuously adapts to stakeholder needs and remains a dynamic, sustainable platform for Massachusetts' AI ecosystem.



Program Management

The project will follow a structured program management framework to ensure delivery of milestones and stakeholder alignment. We will develop and maintain an integrated project plan, reviewed weekly with MassTech stakeholders. Regular steering committee updates will be scheduled to ensure transparency. Progress will be tracked against defined milestones for initiation, infrastructure setup,

dataset curation, synthetic tooling, governance rollout, and public launch. A dedicated program manager will coordinate communications, manage risks, and ensure cross-team alignment.

Team Structure

We have structured the Program Team for strategic oversight, stakeholder engagement, and technical execution, ensuring that the Data Commons is not only successfully launched but also positioned for long-term sustainability, trust, and adoption across the Massachusetts AI ecosystem.

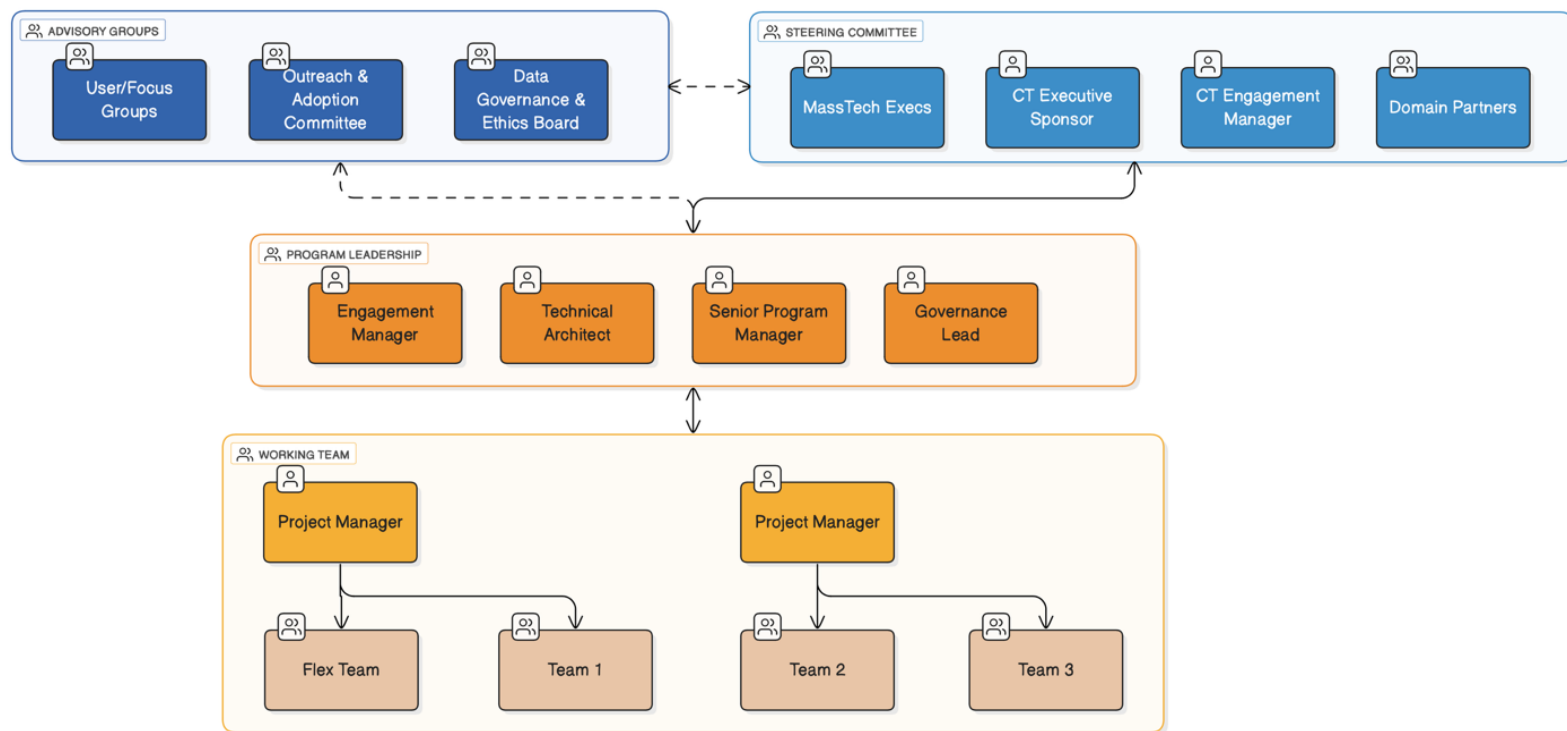


Figure 1 - Program Delivery Structure

Steering Committee

A high-level body composed of senior representatives from the Massachusetts AI Hub, Cleartelligence, and other domain leaders. Provides strategic oversight, ensures alignment with public good and innovation goals, and resolves escalated issues.

- Ensures strategic alignment with state priorities and AI Hub mission.
- Provides accountability and governance, especially around ethical, secure, and fair AI use.
- Builds credibility and stakeholder trust through visible leadership.

Advisory Group

The Advisory Group provides structured feedback, ensures compliance with ethical and legal standards, and guides adoption and engagement strategies. It serves as the voice of the community

and the conscience of the program, ensuring that the Data Commons Collaborative is accessible, equitable, and widely adopted.

- Ensures the Data Commons is user-centered and meets real-world needs.
- Embeds governance, fairness, and privacy principles into decision-making.
- Shapes training and outreach strategies to drive broad and inclusive participation.

Program Leadership

The core management and technical leadership team responsible for day-to-day execution. Includes program manager, technical architect, governance lead, and engagement lead.

- Provides clear accountability for delivery of milestones.
- Ensures program management discipline (planning, reporting, stakeholder coordination).
- Bridges between strategic direction (Steering Committee) and operational execution (Working Team).
- Maintains continuity across all phases

Working Team

This core group of technical, governance, and data specialists executes project deliverables. It includes developers, data curators, security engineers, synthetic data experts, and fairness/bias specialists.

- Provides hands-on expertise to design and build the Data Commons.
- Ensures alignment with FAIR data principles, metadata standards, and governance requirements.
- Enables agile execution with iterative feedback loops from users and leadership.
- Establishes institutional knowledge that can be transitioned to AI Hub staff.

Extended / Flex Team

A bench of specialized experts (domain data experts, cloud architects, synthetic data tool specialists) who can be engaged as needed.

- Provides flexibility and scalability, allowing the team to bring in expertise on demand.
- Reduces cost by avoiding unnecessary full-time staffing.
- Ensures access to niche skills.

Audience Engagement

Engaging external audiences early in the development of the Data Commons Collaborative (DCC) is essential to building a platform that reflects the needs, priorities, and aspirations of the communities it will serve. Early collaboration fosters trust, encourages participation, and ensures that the DCC delivers meaningful value to its diverse stakeholders—including researchers, AI developers, startups, academic institutions, government partners, and nonprofit organizations. We have included a [starting](#)

[framework](#) that we will expand with the Steering Committee to ensure proper alignment to the right audiences and KPIs.

Our team will work closely with MassTech to identify and prioritize the right stakeholders from each of these audiences and to define clear engagement and promotional strategies across the project lifecycle. During development, we will collaborate with user groups to validate concepts, features, and governance models to ensure the platform is intuitive and aligned with real-world use cases.

As we move into launch planning, we will coordinate outreach and promotional activities to build awareness and drive early adoption across the Massachusetts AI ecosystem. Post-launch, we will continue engagement through community forums, feedback channels, and targeted communications designed to sustain participation, promote contributions, and showcase success stories. This structured approach ensures that the DCC is not only technically sound but also community-driven, widely adopted, and positioned for long-term impact.

Technical Approach & Architecture

In this section, we provide technical details on the approach and architecture highlighting how each of the key functionality areas work in unison to deliver a compelling experience for our key audiences. This approach utilizes proven methodologies and technologies. We have included the technical justification for recommended solutions in the [Recommended Services & Platforms](#) section of this document. We welcome the opportunity to discuss and adjust any approaches as needed to align with MassTech's preferred solutions.

Web Experience & Data Management – PortaUS + CKAN

The AI Hub Data Commons (DCC) will leverage [CKAN](#) as the centralized platform for managing all datasets and functionalities. CKAN, with its rich ecosystem of extensions and integrations, will enable the creation of a robust, scalable, and extensible data repository for the DCC.

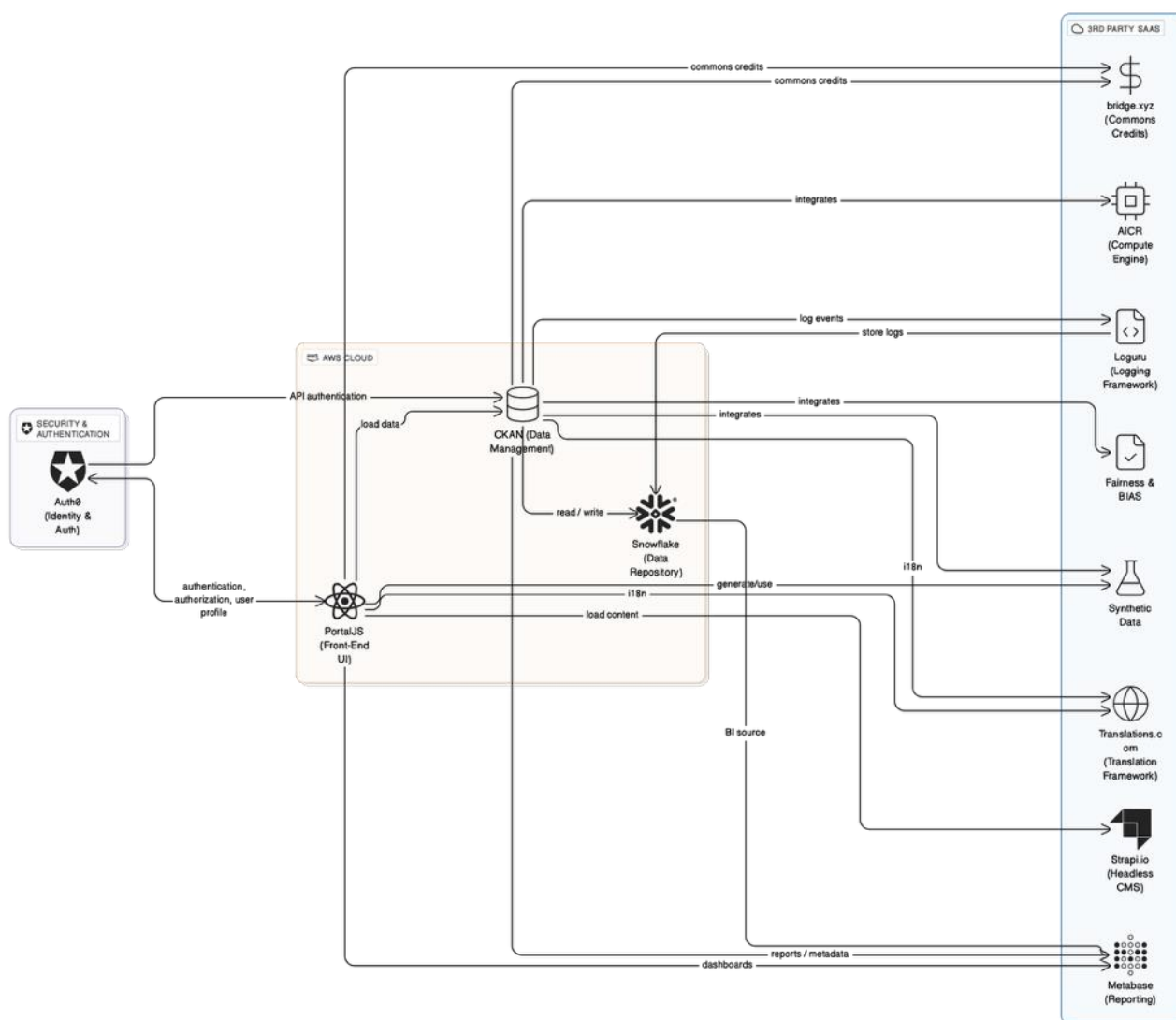


Figure 2 - Technical Architecture Overview

CKAN as the Centralized Hub

CKAN will serve as the core platform to host and manage datasets, providing a seamless experience for users to search, access, and utilize the data. Its flexible architecture allows the hosting of both public datasets from global and local repositories and internally curated datasets, along with support for synthetic data generation tools. CKAN's metadata management framework ensures datasets are findable, accessible, interoperable, and reusable, following the FAIR data principles.

Customizations via the CKAN Extension Framework

The CKAN extension framework will play a key role in customizing the platform to meet the specific needs of the DCC. It will enable seamless integrations with external services, synthetic data tools, and third-party solutions such as AI/ML model platforms and dataset curators. Through CKAN's rich set of extension points, the platform will be customized for role-based access control (RBAC), the integration of synthetic data generation capabilities, and the management of Commons Credits for incentivizing data contributions.

Additionally, custom CKAN extensions will be developed to:

- Facilitate user interactions with external data providers
- Enable automated dataset ingestion pipelines and version control
- Implement ethical data handling policies and compliance with relevant governance frameworks
- Provide APIs for seamless integration with other AI development tools and platforms

CKAN Extensions Targeted for Usage or Development

Extension	Purpose	Source
ckanext-workflow	Defines custom workflow steps and transitions	Open Source
ckanext-envvars	Checks and validates environment variables and applies them to the CKAN configuration	Open Source
ckanext-visualize	Automates dataset visualization	Open Source
ckanext-scheming	Customizes and extends the metadata for the group, organization, and dataset entities. Will be used for custom fields to introduce RBAC for custom DCC capabilities.	Open Source
ckanext-harvest	Provides common harvesting framework for CKAN extensions to manage harvesting sources and jobs	Open Source
ckanext-validation	Provides data validation with the Frictionless framework and validation reports	Open Source
ckanext-approval	Adds role-based approvals to move between workflow steps via custom extension	Custom
ckanext-sso	Provides OpenID connect support for integration with auth0	Open Source
ckan-dcat	Allows CKAN instances to expose their datasets in a semantic RDF format compliant with DCAT specifications, including support for various versions like DCAT-AP v1.1, DCAT-AP v2.1, and DCAT-AP v3	Open Source
ckanext-signal-listener	Listens, logs, and routes actions based on core and custom events fired by CKAN	Custom
ckanext-service-layer	Manages custom actions and 3 rd party calls based on core and custom signals	Custom
ckanext-commons-credits	Manages interactions related to the Commons Credits capabilities	Custom

Figure 3 - CKAN Extensions Table

PortalUS as the Front-End UI

The front-end user interface for the CKAN platform will be powered by [PortalUS](#), a React-based framework. PortalUS will allow us to create modern, dynamic, and user-friendly interfaces for browsing datasets, accessing metadata, and managing user interactions. By using a React framework, PortalUS provides a more responsive and interactive user experience, which is essential for AI/ML researchers, developers, and academics seeking efficient data discovery and exploration.

PortalUS will also enable the following:

- Advanced search functionalities and faceted browsing for datasets
- Previewing datasets and visualizing metadata
- Enabling easy access to synthetic data generation tools and use case integrations
- Ensuring the platform is fully responsive, making it suitable for a variety of devices, including mobile, tablet, and desktop
- Supporting user role management and access control to ensure a secure and compliant data environment
- Ensuring all UI elements will meet WCAG 2.1 AA accessibility standards
- Setting performance benchmarks that will ensure page load times under 2 seconds and search results under 2 seconds for at least 1,000 datasets.

By combining CKAN's powerful backend with PortalUS's dynamic frontend, the DCC platform will be well-positioned for future growth, offering flexibility to expand functionality and incorporate new features as the data ecosystem evolves.

This approach ensures that CKAN is not only a centralized hub for all datasets but also a highly extensible platform that integrates seamlessly with third-party services and data providers. The use of PortalUS as the front-end UI positions the DCC for modular evolution, enabling an agile response to emerging needs and technological advancements.

Web Analytics

To ensure the Data Commons Collaborative (DCC) website is effectively supporting its mission, we will implement [Google Analytics](#) as the primary tool for monitoring and measuring web activity. Google Analytics provides a proven, low-cost, and scalable way to track website traffic, user engagement, and event-driven interactions. By instrumenting the DCC portal with Google Analytics, we will gain visibility into how users discover datasets, navigate the site, and interact with key features such as dataset downloads, synthetic data tools, and Commons Credits functionality. These insights will allow administrators to measure adoption, identify trends, and continuously optimize the user experience in line with the DCC's goals of accessibility, collaboration, and equitable data use.

Google Analytics will also play a role in event tracking and reporting, allowing the platform team to monitor dataset-level engagement, page performance, and access across different user groups. Combined with role-based access controls and audit trails defined in the DCC requirements, this analytics capability will help ensure transparency, support governance, and demonstrate the value of

the DCC to stakeholders. Ultimately, Google Analytics will serve as a foundational measurement layer, supporting both day-to-day operations and long-term evaluation of the platform's impact.

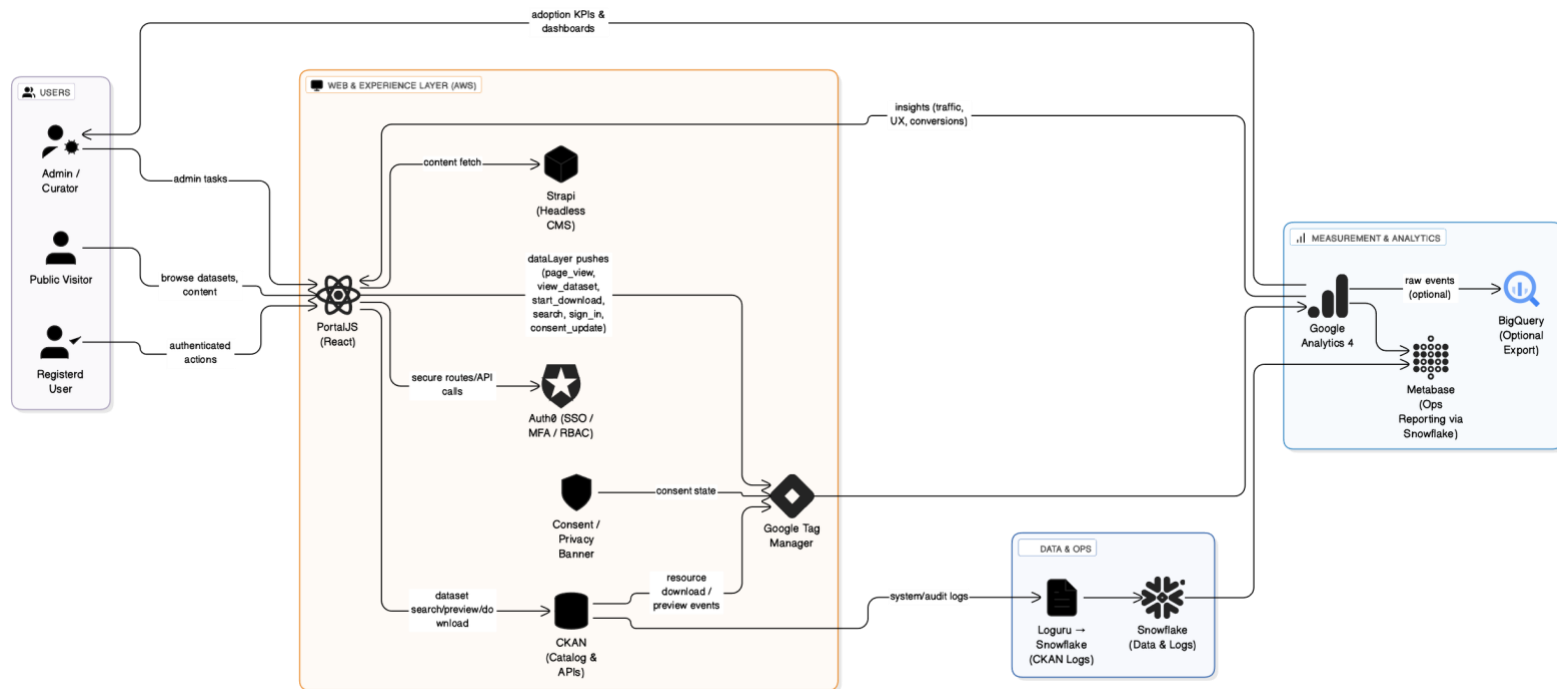


Figure 2 – Google Analytics Integration

CKAN Hosting

There are multiple options available for hosting. For hosting the CKAN platform, we recommend [Amazon Web Services \(AWS\)](#) as the preferred cloud provider. AWS offers the strongest alignment with CKAN's architecture and has a proven track record supporting large-scale government open data portals, including the U.S. federal data.gov platform. Its native services directly map to CKAN's requirements: Amazon RDS (PostgreSQL) or Aurora for the application database, Amazon OpenSearch for search and indexing, and Amazon S3 for object storage of datasets. Amazon also has a deep [partnership with Snowflake](#), our recommended data warehouse. These services are already supported by mature CKAN extensions, enabling a seamless and stable deployment.

In addition, AWS provides the scalability required for the DCC to evolve, with options for containerization on ECS or EKS. Its deep compliance portfolio (FedRAMP High, HIPAA, CJIS, and others) ensures that the platform meets the Commonwealth's governance, security, and privacy requirements. Both Azure and Google Cloud each offer strengths and would serve as viable options, AWS provides the lowest friction for CKAN deployment, the broadest ecosystem of reference implementations, and the most mature integration with CKAN's core storage, search, and database components.

By selecting AWS, the DCC will benefit from a proven, scalable, and secure foundation that minimizes deployment risk while enabling future growth. If desired, we can review and finalize this recommendation during the [Detailed Requirement Gathering](#) phase of this engagement.

Security & Authentication

The DCC will use [Auth0](#) as the centralized identity and access platform, integrated with CKAN through OpenID Connect. Auth0 will enforce RBAC by mapping user claims to defined roles, secure all API endpoints with signed JWTs, and require MFA for sensitive operations. It also provides consent management features to track and enforce user approvals for data sharing. Auth0 is fully compliant with Massachusetts and Federal privacy standards, while supporting both enterprise identity providers and social logins for broad accessibility. This integration establishes a robust and scalable security foundation that balances usability, compliance, and governance for the DCC.

Integration of Auth0 with CKAN

Auth0 will act as the centralized identity provider (IdP), enabling unified authentication and authorization. This approach will utilize the following security standards & solutions for the integration.

- **Protocol:** OpenID Connect (OIDC) with OAuth 2.0.
- **CKAN authn plugin:** Use an OIDC/OAuth2 CKAN extension (e.g., *ckanext-oidc*) configured to Auth0 as the Identity Provider (IdP).
- **Token format:** RS256-signed JWT access tokens (Auth0) with audience(s) for CKAN and any DCC APIs.
- **Claims mapping:** Custom rules/Actions in Auth0 to add role and consent claims; CKAN plugin maps those claims to CKAN org/role permissions on login.
- **Machine-to-Machine (M2M):** Client Credentials flow for backend jobs, ETL, and inter-service calls.
- **Session management:** CKAN uses the IdP session; logout redirects to Auth0 for global sign-out.

Role Based Access Control (RBAC)

CKAN has system and organization-scoped roles. For the DCC we'll standardize a set of business roles, then map to CKAN permissions via claims.

- On login, CKAN's OIDC plugin reads roles (and optionally orgs) claims to assign local CKAN org roles.
- For fine-grained control, we'll add a lightweight CKAN extension to translate additional Auth0 claims (e.g., *dataset_steward:true*, *orgs:[...]*) into CKAN authz checks.
- For APIs, scopes (e.g., *datasets:read*, *datasets:write*, *orgs:admin*) are verified at the resource/action level.

DCC Role (IdP “roles” claim)	Typical Permissions (Business)	CKAN Mapping (examples)
public-visitor	Browse public datasets, docs	Anonymous (no account)
registered-user	Access registered-only content, request access	CKAN user (no org role)
data-contributor	Create/update drafts in assigned orgs	Org Editor; can create/edit own datasets
dataset-steward	Approve/publish datasets, manage metadata quality	Org Admin or custom granular perms via extension
dcc-admin	Global admin, manage orgs, users, extensions	CKAN sysadmin
api-client	Non-interactive API access with limited scopes	No UI role; OAuth scopes only

Figure 4 - Initial Role Mapping Table for RBAC

NOTE: Additional roles and corresponding commissions are expected to be identified and defined during the [Detailed Requirement Gathering](#) phase of the engagement. Custom permissions will be defined using the *ckanext-scheming* and *ckanext-sso* [extensions](#).

Integration Architecture (Auth0 ↔ CKAN ↔ DCC Services)

Auth0 will serve as the DCC's centralized identity and access management layer using OpenID Connect (OIDC) and OAuth 2.0. CKAN will authenticate users through Auth0's Universal Login and receive ID/Access tokens signed with RS256. During login, Auth0 Actions add standardized claims (e.g., sub, email, orgs, roles, scopes, and consent flags) that CKAN consumes to create/update the local user record and apply role mappings. For interactive users, CKAN relies on the IdP session and redirects to Auth0 for single logout. For service-to-service traffic (ETL jobs, internal microservices, and scheduled harvesters), Auth0 issues Access tokens via the Client Credentials flow to callers authorized for narrowly scoped audiences (e.g., `api://dcc-ckan` and `api://dcc-services`).

All CKAN Action API calls and custom DCC service endpoints will validate JWTs against Auth0's JWKS, enforcing issuer, audience, expiry, and scope checks before requests reach the app tier. This architecture externalizes identity concerns from CKAN, concentrates policy in one place (Auth0), and keeps the DCC services clean, consistent, and future-proof.

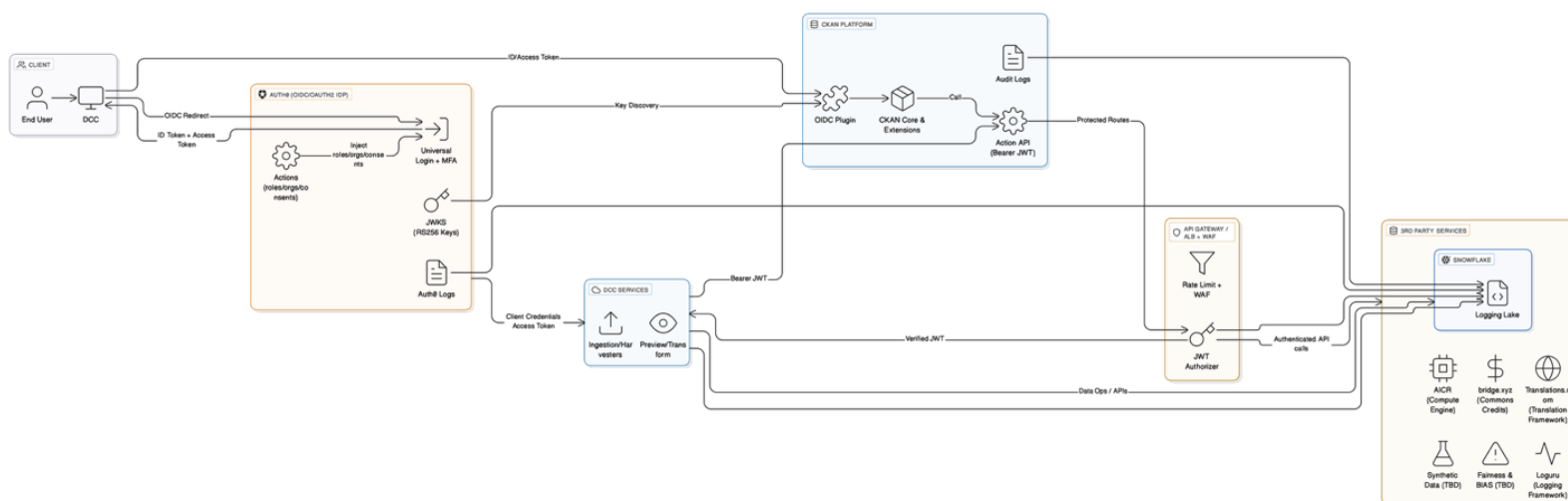


Figure 5 - Auth0 Integration Architecture

RBAC: DCC Role Model & CKAN Mapping

We'll implement role-based access control by mapping Auth0 role/group claims to CKAN's built-in authorization model and (where needed) a thin CKAN plugin for finer policy checks. The core business roles align to the DCC framework as follows:

- Administrators → CKAN sysadmin; full platform configuration and governance.
- Data Developers (Curators/Editors) → CKAN organization Editor/Admin; can create, edit, validate, and publish datasets.
- Data Users (Researchers/Analysts) → CKAN user with read access; can request elevated access to restricted datasets.
- Challenge Participants → Time-boxed role with minimal write and scoped read permissions.
- Auditors / Policy Analysts → Read-only access to datasets, metadata, logs/metrics dashboards as permitted.

- On authentication, CKAN's OIDC extension reads the roles and orgs claims to bind users to organizations and assign CKAN roles. Where CKAN's coarse permissions aren't sufficient, a lightweight extension will interpret additional claims to allow/deny specific actions.

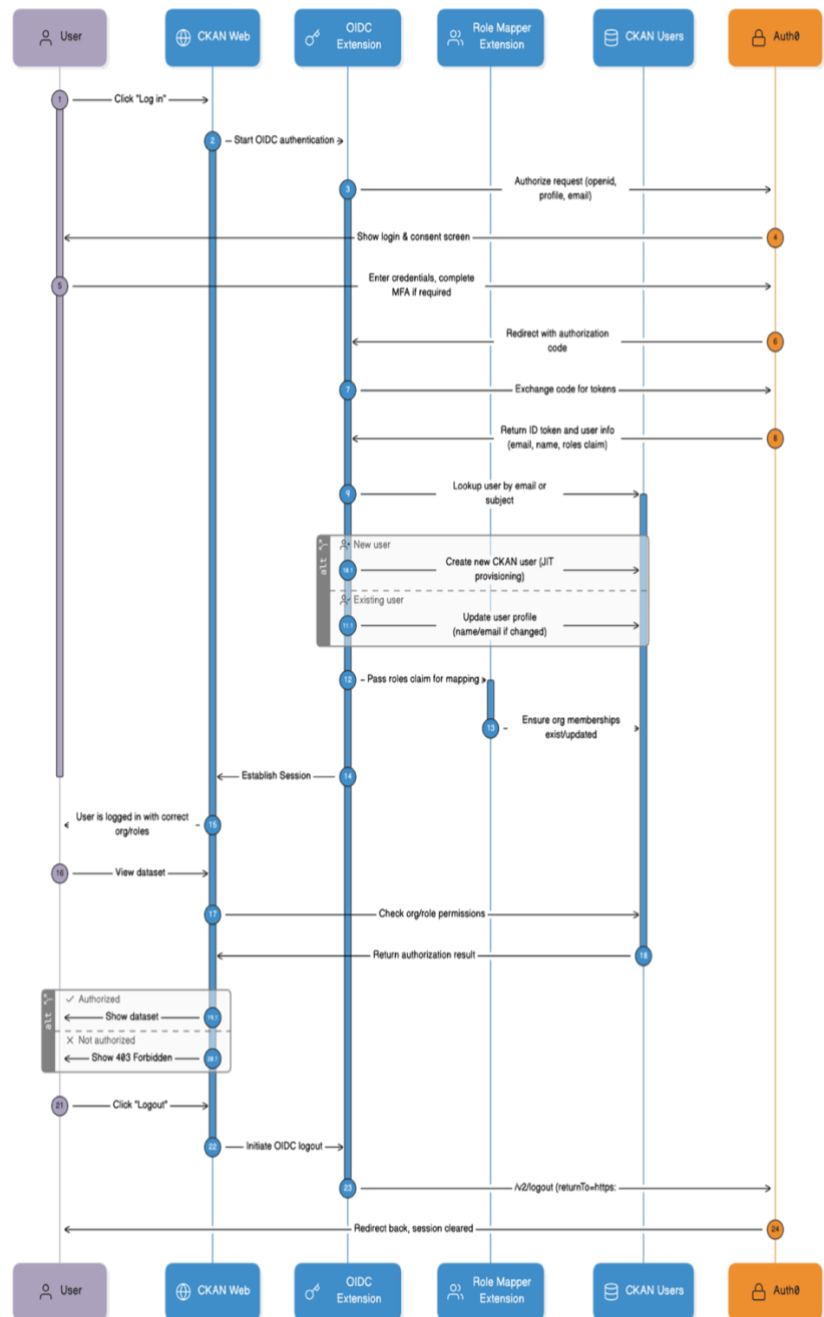


Figure 6 - Auth0 RBAC Mapping

Securing API-Based Endpoints (CKAN Action API + DCC Services)

Every API surface (CKAN's Action API, dataset preview services, ingestion/harvesters, and any DCC microservices) will require Bearer tokens issued by Auth0. Tokens will be verified against Auth0 JWKS (RS256), enforcing iss, aud, exp, and scope. Separate audiences (e.g., api://dcc-ckan vs. api://dcc-services) will prevent token reuse across services. Scopes such as *datasets:read*, *datasets:write*, and *orgs:admin* will gate routes, with CKAN re-checking ownership/org constraints on write operations. APIs may be optionally fronted with AWS API Gateway/ALB JWT authorizers for defense-in-depth. Service accounts using Client Credentials will be scoped to least privilege.

Required Capabilities & Integration Notes

- MFA: Auth0 MFA (WebAuthn, TOTP, SMS, etc.) will be enforced by policy, with CKAN trusting Auth0's assurance.
- Privacy: Auth0 enforces session policies, anomaly detection, and breach checks, minimizing PII shared with CKAN.
- Consent Management: Auth0 Universal Login will capture consent, persist it in app_metadata, and issue as claims checked by CKAN.
- Integration with Other User Stores: Auth0 will support enterprise IdPs (SAML, OIDC, LDAP, Entra, Okta, Ping) to allow for partnership with other parties for login in the future.
- Social Sign-On: Auth0 enables providers (Google, Microsoft, Apple, LinkedIn, etc.), with MFA/risk checks for privileged access.

Configuration Highlights

- Auth0: Configure Applications for CKAN web login and M2M jobs; define APIs (api://dcc-ckan, api://dcc-services) and scopes; Actions inject roles/orgs/consents into tokens; enforce MFA; connect to enterprise/social IdPs; stream logs to SIEM.
 - CKAN: Install OIDC plugin; map claims to CKAN roles/orgs; implement thin authz extension for consent-aware actions; enforce Bearer token use.
- Gateway (optional): Validate JWTs, rate-limit, and add observability controls.

Operational & Security Considerations

Will enforce least privilege, with time-bound elevation for rare admin tasks. Break-glass accounts will require hardware-based factors. JWKS will be cached with short TTL to support rotation. PII remains at Auth0; CKAN receives only minimized claims. Monitoring will correlate Auth0 sub identifiers with CKAN audit logs for full traceability. Runbooks will cover incident response, consent revocation, and user lifecycle management across all IdPs.

Metadata Standards and Interoperability

The Data Commons Collaborative will implement a metadata management framework that fully aligns with internationally recognized standards, ensuring dataset discoverability, interoperability, and long-term reuse. CKAN, as the core cataloging platform, natively supports Dublin Core metadata elements and provides extension frameworks to incorporate additional schemas. Specifically:

- **Dublin Core:** CKAN's core fields (title, description, creator, tags, license, creation date) map directly to Dublin Core elements (e.g., dc:title, dc:description, dc:creator, dc:subject, dc:rights). This ensures compatibility with libraries, archives, and general digital repositories.
- **Schema.org:** Through structured metadata output, CKAN can expose dataset records as *schema:Dataset* objects. Core attributes such as dataset name, description, keywords, publisher, license, and access URLs are aligned with Schema.org fields. This increases visibility in major search engines (Google, Bing, etc.) and enhances machine-readability on the web.
- **DCAT (Data Catalog Vocabulary):** CKAN's *ckanext-dcat* extension enables publishing and harvesting datasets in DCAT and DCAT-AP formats. Each dataset is expressed as a *dcat:Dataset* with related *dcat:Distributionobjects*, ensuring seamless interoperability with external data portals such as Data.gov, the European Data Portal, and other W3C-compliant catalogs.

By implementing these standards, the DCC will ensure conformance with FAIR data principles (Findable, Accessible, Interoperable, Reusable) while maximizing dataset discoverability across academic, government, and industry platforms. The metadata framework will also enable automated harvesting, API-based integrations, and semantic search capabilities.

An initial mapping of CKAN data fields to each of the Metadata standards above is provided in the [Appendix](#). This mapping will be finalized in the [Detailed Requirement Gathering](#) phase of the project.

Common Credits

The Massachusetts AI Hub Data Commons Collaborative (DCC) requires a Commons Credits system that is auditable, transparent, and flexible. The proposed architecture integrates [Bridge.xyz](#) into the DCC through the CKAN extensions framework, ensuring that every action (i.e. Dataset publication, premium data access, or compute job submission) is tied to a credits transaction. This integration leverages Auth0 for identity and authorization, CKAN's native and custom signals for event-driven processing, and Bridge's API for credits issuance, redemption, and settlement.

Architecture at a Glance

The solution layers the Commons Credits system directly within CKAN by creating a dedicated extension (*ckanext-commons-credits*). This extension subscribes to native CKAN signals (e.g., dataset creation, resource download) and emits custom signals that map directly to credits policies. Each credits event is processed by a lightweight service layer that validates policies, calls the Bridge.xyz API, and records the resulting transactions.

Auth0 provides a single source of truth for user authentication and authorization. Each CKAN user is mapped to an Auth0 identity, which in turn is associated with a Bridge account. In this way, credits are securely tied to user or organizational activity, with full traceability from CKAN → Auth0 → Bridge.

Identity & Access

Users authenticate into the DCC via Auth0 Single Sign-On (SSO). Auth0 issues tokens that carry unique user identifiers and role claims (e.g., researcher, curator, admin). The CKAN extension maps these claims to CKAN's RBAC model and to corresponding Bridge account IDs.

When an event occurs that impacts credits, the extension attaches the Auth0 subject ID and Bridge account ID as metadata in the transaction. This ensures that every credits operation is linked to a verified identity, compliant with DCC's security and governance expectations.

CKAN Signals → Credits Events

The credits integration relies on CKAN's signal system to capture activity in real time:

- Native signals capture existing workflow events such as `dataset_created`, `resource_downloaded`, and `user_logged_in`.
- Custom signals extend this framework to represent credit-specific activities, such as `credits.earn.dataset_published` or `credits.spend.compute_job_started`.
- Each signal carries a structured payload with user identifiers, dataset/resource IDs, the credit delta, and purpose tags (e.g., "data_contribution," "premium_access"). These signals are routed through the extension's service layer and recorded both locally and in the Bridge ledger.

Extension Design (*ckanext-commons-credits*)

The Commons Credits extension is designed to be modular, following CKAN plugin conventions. It provides:

- Signal subscriptions that map CKAN actions to credit rules.
- REST endpoints for balance queries, transaction history, and admin functions.
- Policy enforcement hooks to check balances before allowing premium actions.
- UI enhancements (e.g., a balance HUD, dataset contribution rewards, premium download cost indicators).

Internally, the extension hosts a service layer that implements credits policies (rules for earning/spending), calls the Bridge API with secure credentials, and maintains a local mirror of transactions for fast UI rendering.

Event → Action Flows

A typical earn credits workflow begins when a user publishes a curated dataset. CKAN emits *dataset_created*, which triggers a custom signal (*credits.earn.dataset_published*). The service layer then issues credits through Bridge, records the transaction, and updates the user's balance in the UI.

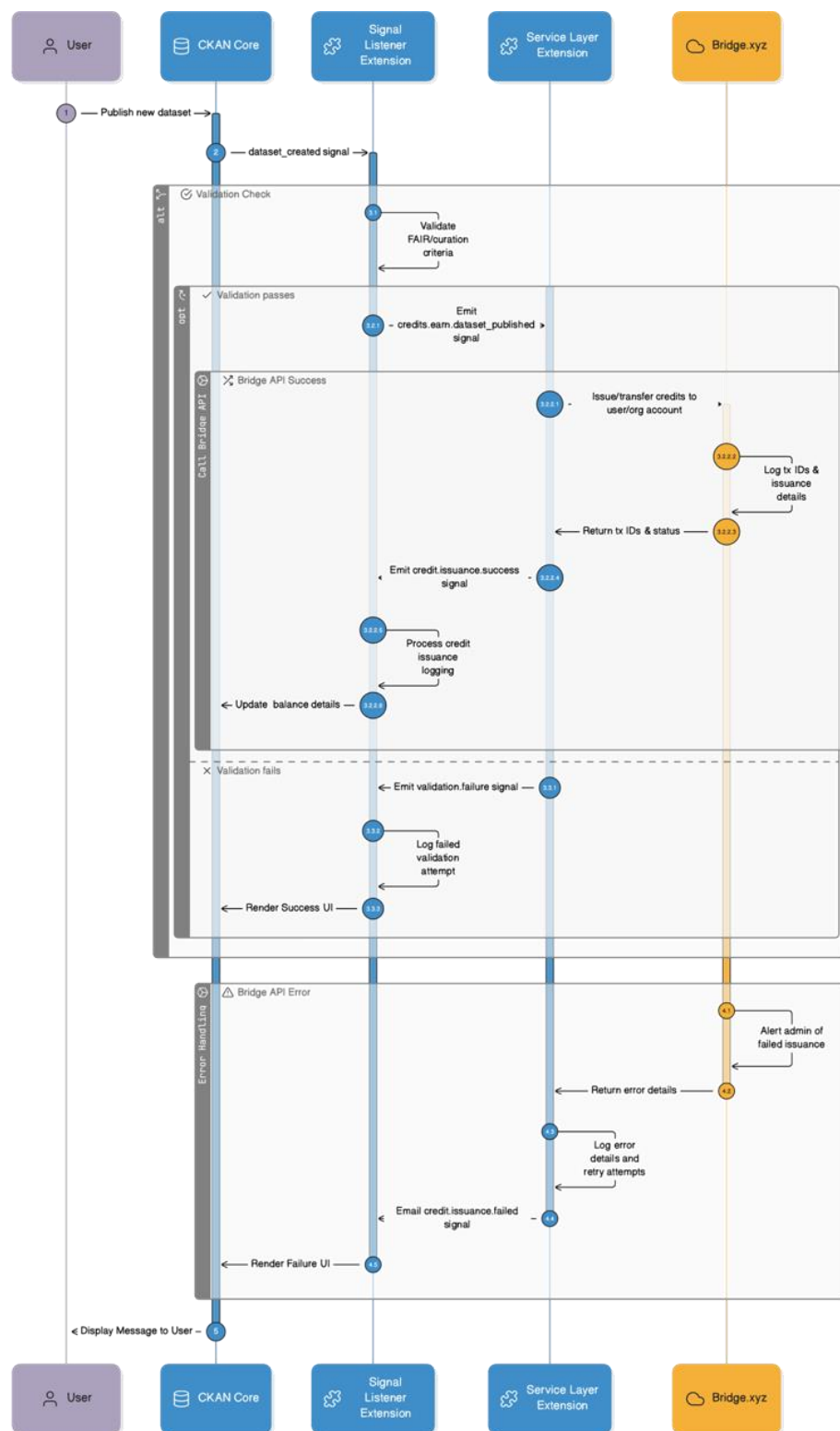


Figure 7 - Credits Earned for Dataset Upload

For spending credits, such as downloading a premium dataset, the extension intercepts the download request, checks the balance via Bridge, and only grants access if sufficient credits exist. Credits are redeemed, the ledger updated, and access provided via a secure URL.

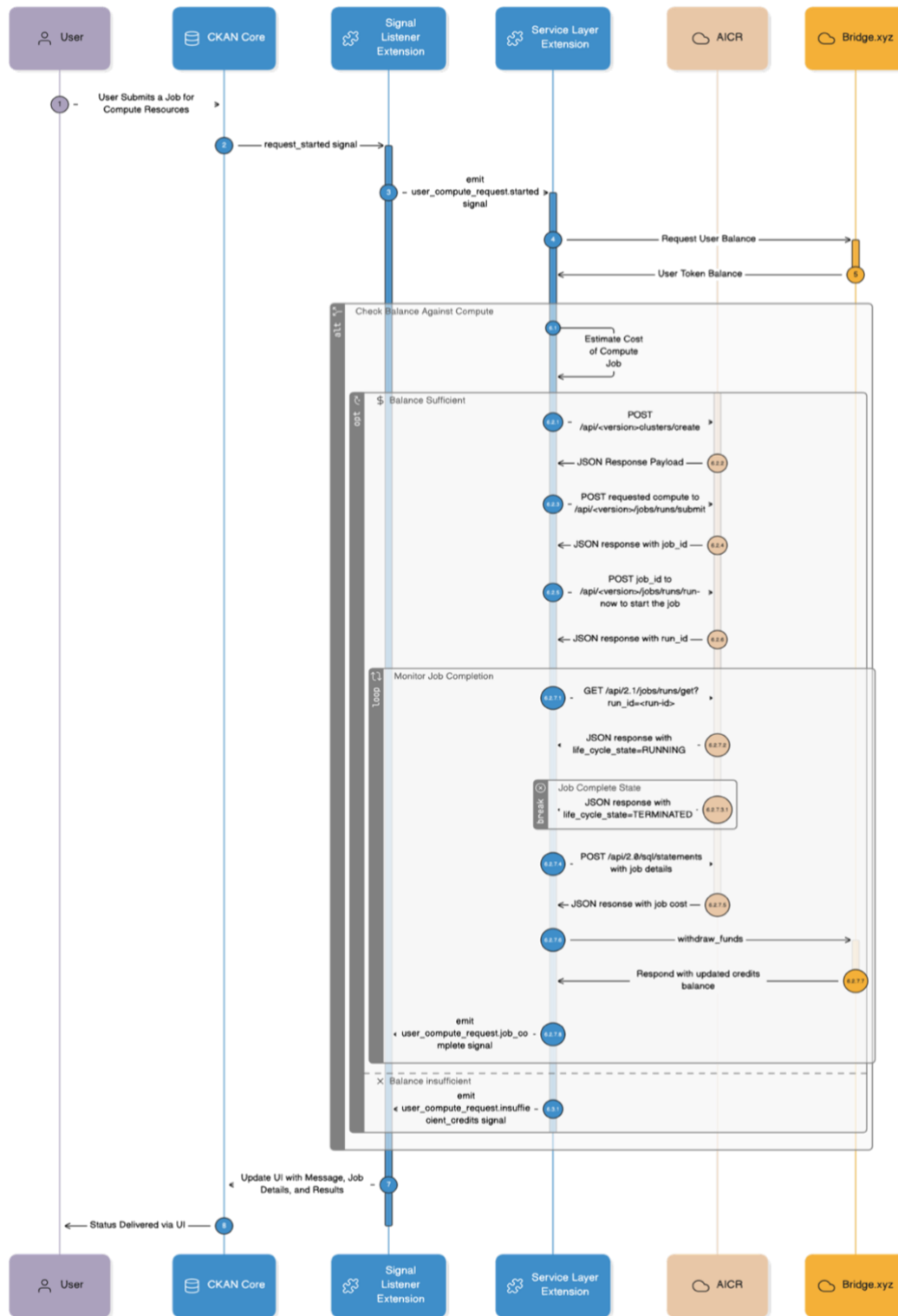


Figure 8 - Spend Credits on Compute

Webhooks & Reliability

To maintain ledger consistency, Bridge webhooks notify the extension of completed transactions, reversals, or refunds. Failed API calls are retried with idempotency keys to prevent duplication. This ensures that the Commons Credits ledger remains accurate, auditable, and resilient.

Security & Compliance

The integration follows a least-privilege model. Bridge API calls are made with service account credentials, while user identities are carried as metadata. Secrets are stored in a vault, and sensitive identifiers are pseudonymized. Audit trails record every credit event with full metadata, meeting the DCC's governance and security requirements.

Configuration & DevOps

- **ckan.ini** entries for Auth0 (issuer, client_id, audience), Bridge (base URL, creds), and policy tables.
- **Feature Flags:** enable/disable categories (downloads, compute, grants) per phase.
- **Observability:** metrics (balances, issuance, redemptions, failures), alerts on policy breaches or spend spikes.
- **Testing:** contract tests against Bridge sandbox, signal unit tests, end-to-end flows (publish → earn → spend → reverse).

Warehousing & Hosting

The Massachusetts AI Hub Data Commons Collaborative (DCC) requires a secure, scalable, and widely adopted platform for hosting and managing data. Our recommendation is [Snowflake](#), the state's existing enterprise data warehouse of choice. Snowflake's cloud-native architecture, strong security posture, and ease of integration make it uniquely aligned with the Commonwealth's requirements. Its widespread use across Massachusetts agencies ensures operational familiarity, streamlined procurement, and reduced onboarding time. While [Databricks](#) was evaluated and is a viable option, particularly given its strengths in advanced analytics and AI workflows, the Commonwealth's existing experience and investments on the Snowflake platform led us to use it in our solution approach. The approach described in this section would not change if MassTech decides to utilize Databricks over Snowflake.

The integration of CKAN with Snowflake is designed to achieve three primary objectives. First, Snowflake will provide the authoritative platform for dataset hosting, with CKAN acting as the data catalog and metadata manager. This ensures that CKAN serves as the entry point for dataset discovery, curation, and access control, while Snowflake provides secure and scalable storage and query capabilities. Second, Snowflake will act as the centralized repository for operational logging and auditing. CKAN will generate logs of user activity, dataset ingestion, and publishing events, while additional API-level interactions with third-party services such as Auth0 and bridge.xyz will also be captured. Finally, Snowflake will expose curated and structured views for downstream reporting and analytics, enabling tools like Metabase to deliver insights into usage, governance, and synthetic data workflows.

Reference Architecture

The architecture supporting this integration places CKAN at the center of cataloging and access while positioning Snowflake as the back-end data warehouse. When datasets are uploaded or registered through CKAN, the system records metadata and enforces governance policies. Raw files are placed into cloud object storage, where they are staged for ingestion into Snowflake. Snowpipe and related pipeline services then process and load these files into structured tables, creating curated schemas that reflect the dataset's intended structure and governance tags. CKAN extensions are used to manage this process, triggering ingestion jobs and synchronizing metadata between the two systems. Snowflake, in turn, becomes the single source of truth for hosted datasets, logs, and analytic-ready views, while CKAN remains the point of interaction for users and administrators.

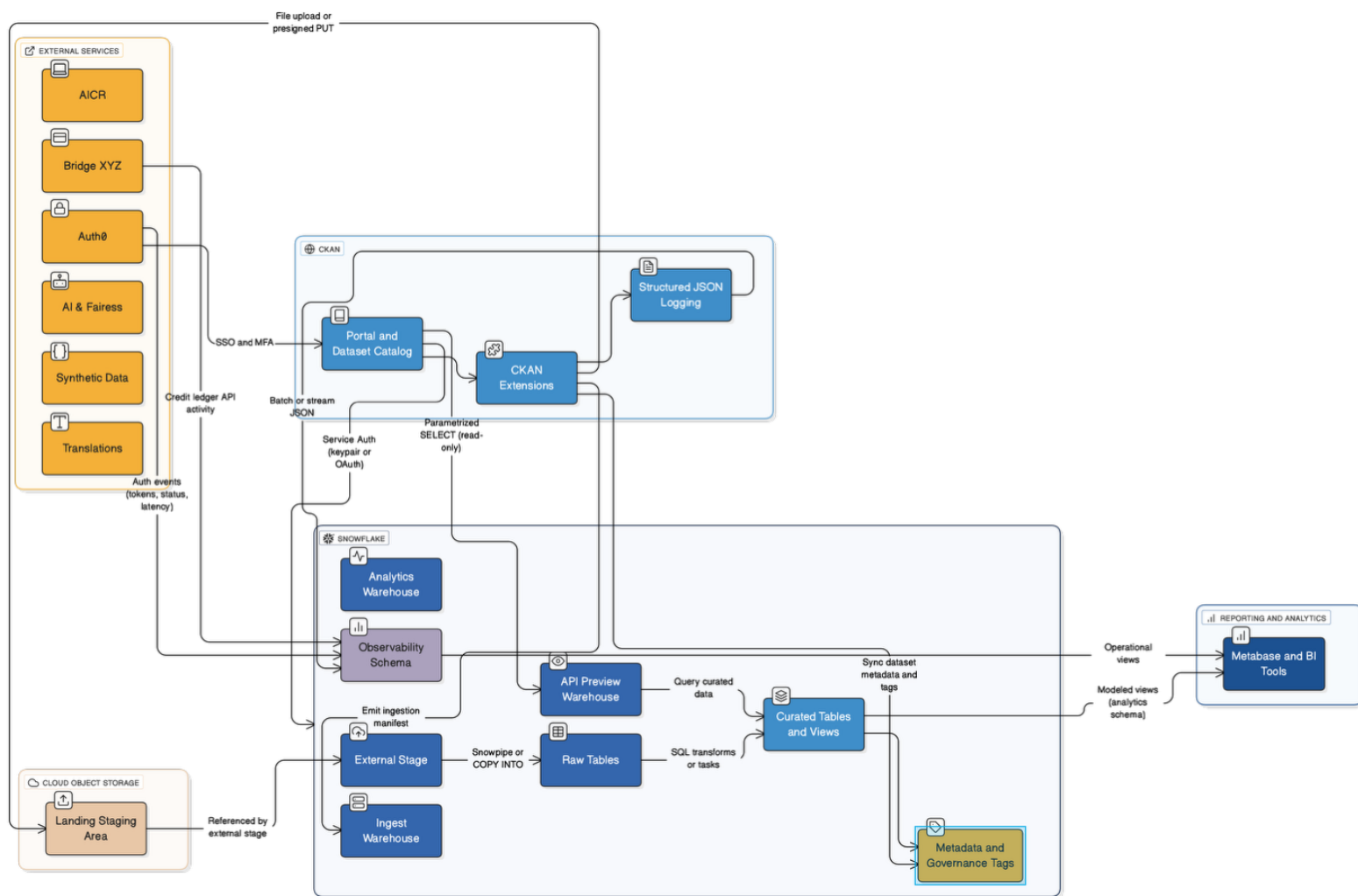


Figure 9 - Snowflake Integration Architecture

Dataset Hosting Flow

When a dataset is introduced into the system through CKAN, the platform validates its format, captures metadata, and generates an ingestion manifest that defines its schema, ownership, and sensitivity attributes. Files are landed in cloud object storage and referenced in Snowflake via external

stages. Snowpipe then loads these datasets into raw tables, followed by transformations that normalize and curate the data. Curated tables may include version control through timestamped tables or Snowflake’s Time Travel feature. CKAN records each version as a dataset revision, maintaining transparency between catalog entries and Snowflake objects. When end users preview data in CKAN, the request is routed through parameterized queries against Snowflake, ensuring that CKAN’s interface delivers authoritative results directly from the warehouse.

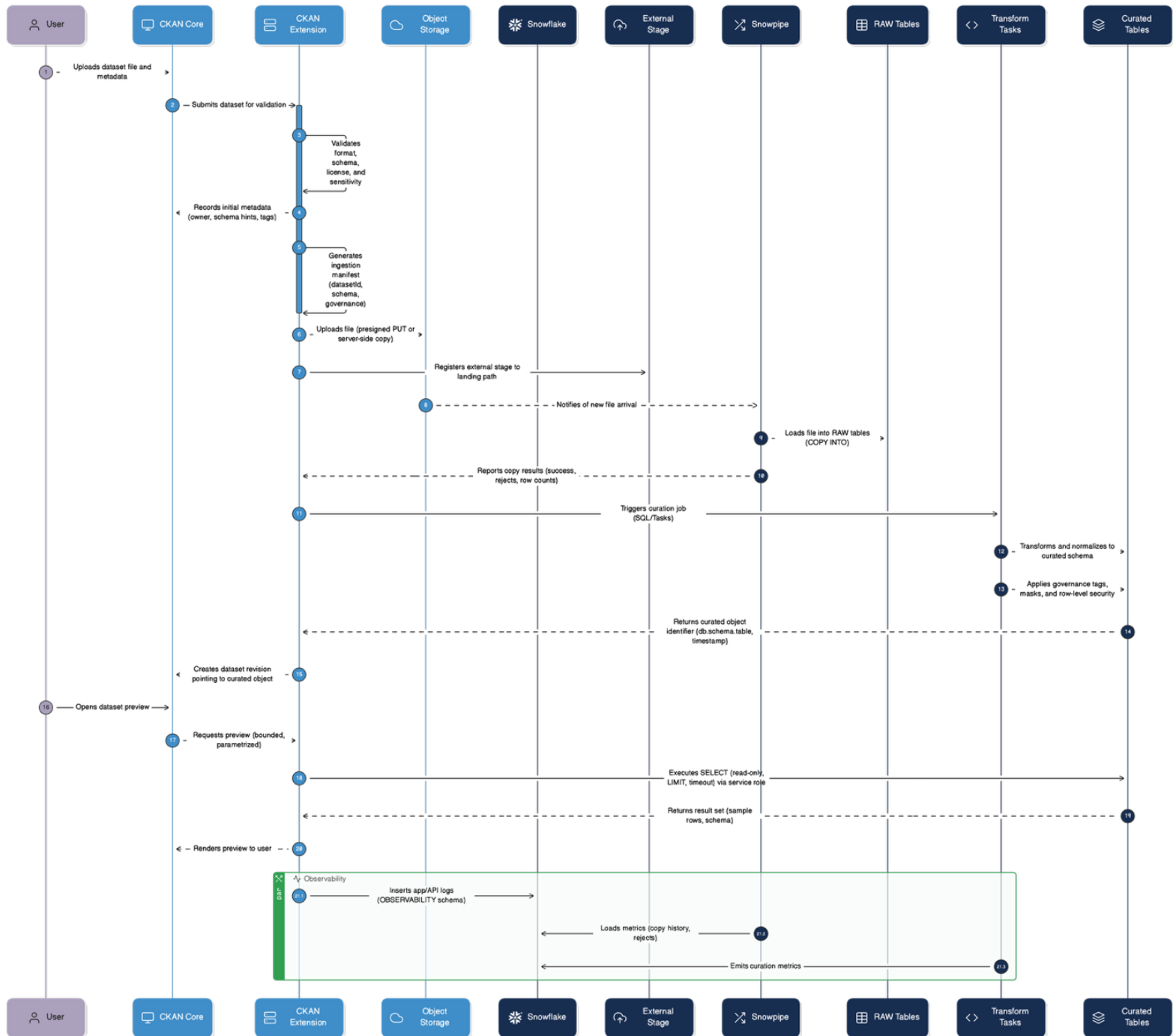


Figure 10 - Dataset Upload with Snowflake Storage

Security and Access Control

Security within this integration is enforced through a layered model. CKAN leverages Auth0 for user authentication, including MFA and role-based access, while Snowflake enforces corresponding roles, row-level security, and masking policies to protect sensitive data. A small set of service roles is created in Snowflake, such as *ingestor*, *preview*, and *reporting* roles, each scoped with the principle of

least privilege. All traffic between CKAN and Snowflake is encrypted in transit, while Snowflake ensures encryption at rest. Dataset-level governance tags originating in CKAN are mirrored into Snowflake's metadata, ensuring consistent policy enforcement. This dual enforcement mechanism ensures compliance with Massachusetts and federal privacy requirements while delivering strong protections for hosted and sensitive datasets.

Operational Logging in Snowflake

Snowflake also provides the central repository for all operational and API-level logs. CKAN emits structured JSON logs that capture dataset-level events, authentication, and errors, which are ingested into Snowflake either through batch ingestion from object storage or through direct API pipelines. In addition to CKAN logs, API-level logs from Auth0 authentication events and bridge.xyz credit transactions are also persisted. These logs are normalized into dedicated schemas within Snowflake, where tables are designed to support governance, monitoring, and auditing. This consolidated logging environment ensures that compliance requirements, including retention and auditability, are satisfied and that administrators can trace all system and user activity across the platform.

Reporting and Analytics

For reporting, Snowflake provides curated views that integrate CKAN metadata, usage logs, and dataset activity. Read-only roles are provisioned for tools like Metabase, which connect directly to Snowflake to generate dashboards and reports. These views capture dataset catalog information, ingestion health, API activity, and governance audits. Dashboards built on this data provide insight into key performance indicators such as ingestion success rates, dataset popularity, sensitive dataset access, and Commons Credit activity. By using Snowflake as the reporting backbone, the DCC ensures that analytics are based on a consistent, authoritative set of data and logs.

Performance, Cost, and Reliability

The integration is designed to balance performance and cost efficiency. Dedicated Snowflake virtual warehouses handle ingestion, preview queries, and reporting workloads, each tuned for its purpose. Auto-suspend and auto-resume capabilities limit costs during idle periods; while caching and materialized views optimize frequent queries. Data pipelines include retry mechanisms, dead-letter queues, and validation checks to ensure reliability and resilience. Versioning and lineage capabilities are preserved through Snowflake's built-in Time Travel and metadata tagging features, providing reproducibility and auditability of hosted datasets.

Configuration and DevOps

From an operational standpoint, CKAN manages credentials for Snowflake through environment variables or secret management tools, ensuring that sensitive keys are not exposed. Connectivity between CKAN and Snowflake is secured through private network links or IP allowlisting. Separate environments for development, testing, and production are maintained, with schema and policy definitions managed as code and promoted through CI/CD pipelines. This approach ensures that integrations remain consistent, secure, and auditable across environments.

In summary, the integration of CKAN and Snowflake creates a clear division of responsibility: CKAN acts as the authoritative catalog and user interface, while Snowflake becomes the authoritative store for datasets and logs. Together, they establish a secure, scalable, and governable architecture that supports dataset hosting, centralized logging, and robust reporting, all while aligning with Massachusetts' standards for data security, compliance, and innovation.

Logging & Centralized Observability

The objectives of the logging approach are to:

- 1) Provide end-to-end visibility across CKAN and the broader Data Commons Collaborative stack
- 2) Ensure that all logs are standardized and structured for reliable analytics, auditing, and troubleshooting
- 3) Centralize both first- and third-party events in Snowflake. This unified logging and observability strategy will support comprehensive reporting, enhance compliance with security and governance requirements, and enable stakeholders to monitor performance, trace issues, and generate insights through centralized dashboards and audit trails.

Scope of Logging

The scope of logging encompasses all major components of the platform.

- Within CKAN, logs will capture API requests, user interface actions, background jobs, extension lifecycle events, dataset operations, and validation errors.
- PortalUS may optionally generate client telemetry such as page load times, API performance metrics, and captured errors, which can be routed through a server-side collector.
- Strapi will produce logs related to content creation, updates, and publishing workflows, while Auth0 will contribute authentication and audit events covering logins, MFA prompts, and consent management.
- The AICR compute engine will log job submissions, execution status, errors, and runtime metrics, and Snowflake itself will generate access and query history that can be used for lineage and operational monitoring.
- Metabase will add metadata on query executions and dashboard usage, and external integrations such as Translations.com and bridge.xyz will stream webhook-based events.

Together, this scope ensures that both system-level and business-level activities are observable across the entire DCC ecosystem.

CKAN Application Logging (Loguru)

Within CKAN, the [Loguru](#) extension will serve as the centralized logging framework, replacing or augmenting the platform's default logging. Logs will be emitted in structured JSON line format, ensuring consistent parsing and downstream analysis. Each log entry will be enriched with contextual fields such as trace and correlation IDs, user identifiers (hashed where required), dataset or resource

identifiers, endpoint details, response status, and request duration. Standard metadata including component name, service version, and environment will also be included.

Log levels will be managed to balance observability and performance: debug events will be used in development only, while info-level logs will serve as the default in production, with warnings, errors, and critical messages escalated accordingly. To prevent accidental disclosure of sensitive data, personally identifiable information will be filtered or hashed, and secrets such as tokens will be redacted before transmission. Logs will be written to local files with rotation and retention policies while also being streamed to a log forwarder for ingestion into Snowflake, where they will be stored and joined with logs from other systems for unified reporting and compliance monitoring.

Event Taxonomy & Schema

To ensure consistency across all logged events, the platform will implement a unified taxonomy and schema. Events will be categorized into key types such as application logs, audit logs, job events, webhook events, access logs, and metrics, with each type designed to capture both technical and business context. Every log entry will include a standardized set of required fields, including event time in UTC ISO-8601 format, event type, severity level, component, descriptive message, trace and correlation identifiers, source host, and environment. Additional domain-specific fields will be included depending on the event type—for example, API events will contain HTTP method, path, status, duration, and response size, while audit events will include hashed user and organization identifiers, dataset identifiers, and the action performed. Compute jobs will capture job IDs, state transitions, runtime, and exit codes, while webhook events will include provider details and payload references. To support long-term governance, schemas will be versioned, maintained in a catalog, and validated through CI processes to enforce backward-compatible changes.

Transport & Ingestion

Log transport and ingestion will be managed through a resilient agent-based architecture. A lightweight forwarder such as Vector or Fluent Bit will collect events from multiple sources, including CKAN's Loguru stream, web server access logs, and application stdout or file-based logs from Strapi and PortaUS. These agents will normalize event formats, enrich logs with Kubernetes or ECS metadata, and apply redaction rules before delivery. From there, events will be routed to cloud storage for Snowpipe-based ingestion into Snowflake, where notifications will trigger automatic loading into raw landing tables.

For use cases requiring lower latency, Snowpipe Streaming will be considered to deliver near-real-time observability. Third-party events such as Auth0 authentication logs will be streamed through their APIs or native log connectors into the same pipeline, while webhook-based integrations from providers like Translations.com and bridge.xyz will be received through an ingestion microservice that validates signatures, applies replay protection, and routes the events into the logging pipeline. To ensure resilience, the transport layer will support retries with exponential backoff, maintain at-least-once delivery semantics, and use idempotent merge operations in Snowflake to handle duplicates or late-arriving events.

Snowflake Modeling & Storage

Within Snowflake, logs will be staged into raw landing tables stored as JSON, partitioned by date and hour with configurable retention policies. From these raw tables, transformation processes will flatten and type logs into domain-specific fact tables, such as application events, audit records, or compute job histories, supported by dimension tables for users, datasets, and services. This structured model will enable analytics for reliability, security, and product adoption. Deduplication logic will ensure idempotency, while late-arriving events will be processed through streams and tasks with appropriate watermarking. Derived marts will support reporting use cases such as operational error budgets, API latency metrics, audit trails, and adoption dashboards, all accessible via Metabase for visualization.

Monitoring, Alerting & SLOs

Monitoring and alerting will be configured to provide real-time visibility into system health and compliance. Alerts will be triggered when ingestion pipelines stall, error rates spike, or thresholds for failed logins, MFA errors, or compute job failures are exceeded. Metabase dashboards will be used to visualize API latency, error heatmaps, ingestion freshness, and adoption metrics. Service Level Objectives (SLOs) will be defined for key areas, such as API availability at 99.9 percent, ingestion freshness within five minutes for priority streams, and p95 latency targets for critical endpoints. Error budgets will be tracked and reviewed periodically to prioritize remediation and ensure reliability commitments are met.

Log Retention & Auditability

By utilizing Snowflake as the centralized repository for logging data, we gain native capabilities to achieve the requirement to maintain 3 years of auditable logs. To reduce log storage costs while maintaining accessibility, Snowflake data can be managed across a tiered storage lifecycle leveraging both Snowflake and AWS S3:

1. **Hot Storage (Snowflake Native)** - Keep the most recent 12 months of logs in Snowflake's native tables for fast, frequent queries.
2. **Warm Storage (S3 Intelligent-Tiering)** - Data older than 1 year is automatically offloaded from Snowflake into an S3 bucket with Intelligent-Tiering, which optimizes costs based on access patterns. External tables in Snowflake allow seamless querying across Snowflake + S3.
3. **Cold Storage (Glacier Instant Retrieval)** - After 2 years, data transitions into Glacier Instant Retrieval, minimizing storage costs while remaining queryable (with slightly slower performance).
4. **Expiration / Deletion** - After 3 years, data is automatically expired and deleted, preventing unnecessary storage costs.

Automation is handled via Snowflake Tasks and Procedures to unload data daily and manage retention seamlessly. Views provide unified query access to both hot and cold storage layers, maintaining transparency for users.

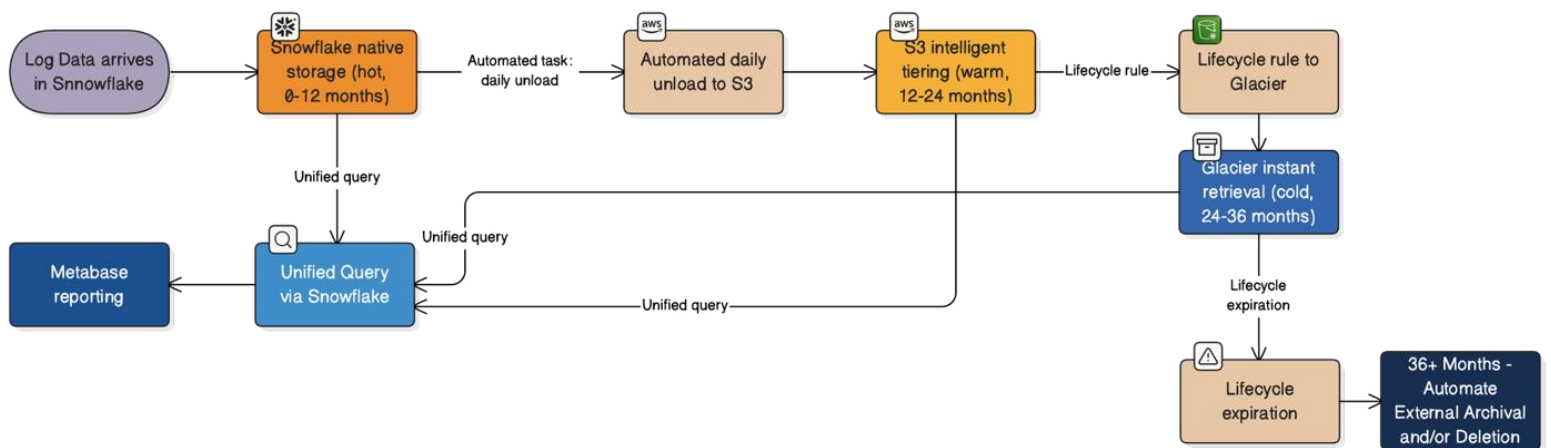


Figure 11 - Log Archival & Retention

Reporting & Dashboards

[Metabase](#) will serve as the reporting and analytics layer within the DCC platform, tightly integrated with both Snowflake and CKAN. Snowflake will remain the primary repository for curated and versioned datasets. Metabase will connect directly to Snowflake through its optimized connectors, ensuring secure, performant access to raw and curated data. This approach guarantees that reporting is always based on authoritative data, with full lineage and governance preserved.

Within the portal architecture, Metabase will provide charts, dashboards, alerts, and analytics components that are embedded directly into CKAN. By leveraging the SaaS Pro tier of Metabase, we can take advantage of React-based embedding rather than relying on iframe embedding. This modern approach ensures that visualizations feel native to the DCC portal, maintain consistent styling, and respect user permissions defined through CKAN and the identity management framework. Metabase will support three main integration patterns:

1. **End-User Visualizations in CKAN:** Researchers, startups, and other DCC users browsing datasets in CKAN will be able to view interactive charts and dashboards powered by Metabase. These embedded visualizations provide a quick way to assess data without requiring downloads or external tools, enhancing usability and adoption.
2. **Administrative Dashboards:** Platform administrators will access Metabase dashboards to monitor dataset usage, track Commons Credits activity, and review system metrics. This supports the governance and auditing requirements of the DCC, with reporting aligned to FAIR data principles and role-based access controls.
3. **Automated Alerts and Subscriptions:** Users and administrators will be able to configure alerts in Metabase that trigger when a dataset, dashboard metric, or Commons Credits balance changes. Alerts can be delivered via email, Slack, Teams, or webhook integrations, providing proactive insight into critical events without requiring constant manual monitoring.
4. **Fairness Dashboards & Reports:** In addition to system usage dashboards, Metabase will power fairness and bias dashboards that visualize bias detection metrics (e.g., disparate impact ratios, fairness scores) from integrated tools.

This architecture ensures that Snowflake provides the single source of truth for data, CKAN serves as the catalog and access layer, and Metabase delivers embedded analytics and proactive alerts seamlessly within that ecosystem.

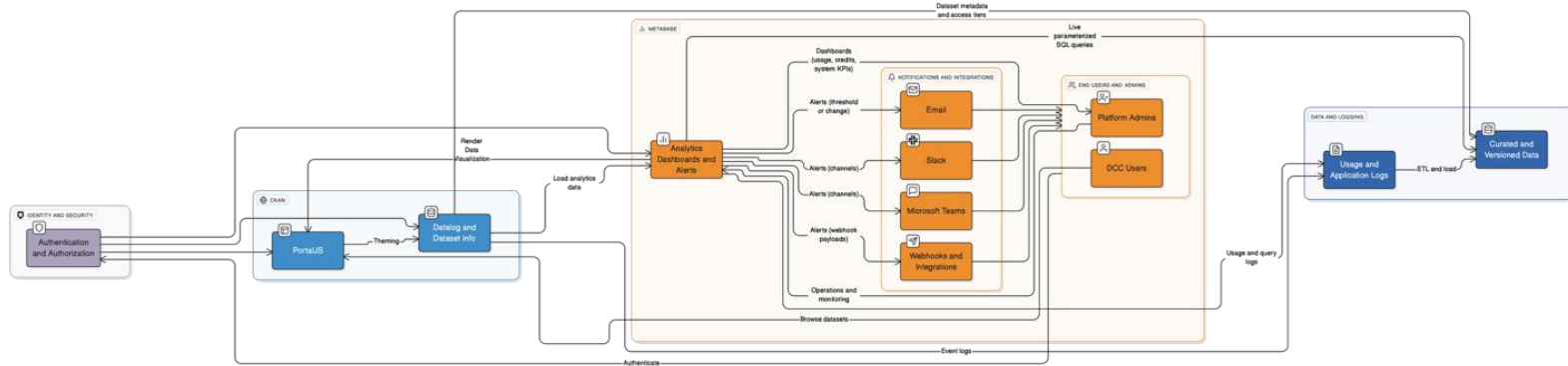


Figure 12 - Metabase Integration Architecture

Dataset Curation, Integration, & Governance

At the heart of the Data Commons Collaborative (DCC) lies CKAN, which acts as the system of record for all datasets and their metadata. Every dataset, whether it is hosted directly within the Commons or referenced from an external source, has a canonical entry in CKAN with structured metadata, classifications, ownership, and lifecycle attributes. This ensures consistent governance across the ecosystem. The solution is built around separation of concerns: CKAN manages cataloging, policy, and metadata; Snowflake provides the scalable repository for raw and curated datasets; PortalJS powers the modern front-end experience; Auth0 enforces identity and multi-factor authentication; and Loguru with Snowflake consolidates audit trails. Metabase and Strapi extend the platform with governed analytics and content management. Importantly, workflows for intake, review, governance, discoverability, and retention apply equally to both hosted and external data, ensuring parity and compliance.

Dataset Categories & Storage Patterns

Datasets entering the Commons fall into two categories: hosted and external. For hosted datasets, CKAN manages uploads into AWS object storage, where files are encrypted and governed by bucket policies. CKAN resource entries then reference these objects, and large files are handled through multipart uploads. Ingestion manifests produced by CKAN trigger Snowpipe to load raw data into Snowflake, where dbt or equivalent tools transform them into curated schemas with data contracts and validation tests. For previews, CKAN resource views call a secure proxy that submits parameterized queries to Snowflake, enforcing row-level filters aligned with user permissions.

Externally hosted datasets are governed without replication. CKAN registers metadata records for URLs, APIs, or remote resources such as ArcGIS layers or external S3 buckets. Each registration captures licensing, consent flags, and provenance. The system continuously verifies availability through link checking and, when policy permits, can mirror or cache content for stability. Where replication is prohibited, CKAN maintains timestamps of last verification and optionally stores hashes or signatures for integrity. For APIs, CKAN retains endpoint details, authentication methods, and service limits, while a secure proxy can enforce usage policy and collect observability metrics without copying data in-house.

Dataset Curation, Integration, & Governance

Dataset curation will follow clear inclusion criteria—datasets must be high-value for AI/ML research, well-documented, and either openly licensed or legally shareable. Provenance metadata will record dataset origin, processing steps, and version lineage. Collaborative workflows will include multi-step review and approval processes involving dataset contributors, stewards, and administrators.

Metadata, Schemas, and Validation

Metadata management is foundational to trust. CKAN's Organizations, Groups, Tags, and Custom Fields (implemented via the scheming extension) define required attributes such as ownership, stewardship, license, sensitivity, retention, and lineage. Metadata is published using standards like DCAT/DCAT-AP for interoperability, and where relevant, ISO or Frictionless Data schemas for technical detail. On ingestion, files and APIs undergo validation checks that include format detection, schema conformance, record counts, and PII scanning for sensitive classes. These validations ensure quality and compliance before publication. To support accessibility across Massachusetts' diverse communities, field-level multilingual metadata and UI strings are integrated through Translations.com.

Access Control, Consent, and Classifications

Identity management and access enforcement are powered by Auth0, which integrates single sign-on, MFA, and social logins where permitted. CKAN's native role-based access model is extended with custom roles such as Data Steward and Compliance Reviewer. Every dataset is assigned a classification (i.e. Open, Restricted, Confidential, or Protected PII) driving access rules. Consent metadata, including legal basis, purpose, and expiry, is stored with each dataset and is enforced through workflow gates before publication. Hosted data enforces row- and column-level restrictions in Snowflake using role-based policies, while external APIs rely on either the source's own access controls or the Commons' secure proxy.

Governance Workflow (Hosted & External)

The lifecycle of a dataset follows a consistent workflow. It begins with intake, where submitters provide metadata and files or references, and automated validation is applied. In the steward review phase, a Data Steward examines schema, classification, licensing, and consent, resolving quality or lineage issues. Next, the compliance gate checks legal terms, retention requirements, and export restrictions. Once approved, the dataset is moved to published state and becomes available within the Commons under the proper visibility rules. CKAN records all approvals in its activity stream, with logs persisted in Snowflake for audit. Updates follow the same workflow, with CKAN maintaining revision history. Finally, datasets may be archived or de-published when they expire, are withdrawn, or no longer meet policy, ensuring lifecycle discipline.

Data Quality & Observability

Automated and manual checks establish confidence in data quality. Integrity checks, schema validation, and statistical profiling are applied at intake. Outlier detection and quality scoring can surface issues to stewards and end-users. External resources undergo regular link health monitoring,

with broken or degraded links triggering notifications and issues. CKAN tracks lineage from raw to curated states using metadata and job identifiers, and where required, integrates with lineage catalogs like DataHub. All catalog events, approvals, and access decisions are logged via Loguru and centralized in Snowflake, ensuring observability across the system.

Discovery & UX (PortalUS + CKAN)

The user experience is delivered through PortalUS, which exposes the CKAN catalog in a modern, React-based interface. Users can search datasets by organization, topic, geography, license, classification, steward, recency, or quality score. Each dataset page includes purpose, lineage diagrams, owner contacts, and license terms. Where applicable, “how to use” examples in SQL or API calls are provided. Embedded Metabase charts and CKAN resource previews offer immediate insights, with visibility enforced by Auth0 and CKAN roles. This ensures users see only the data they are entitled to while benefiting from intuitive navigation.

Interoperability & Exchange

The Commons is designed to operate in a federated ecosystem. CKAN harvesters bring in external catalogs that meet policy, including DCAT, Socrata, ArcGIS, and CKAN-to-CKAN feeds. These harvested datasets are subjected to the same governance workflows before being published. Outbound interoperability is achieved through CKAN REST APIs, DataStore APIs, and DCAT feeds segmented by classification or organization. Hosted downloads use signed URLs to secure access, and API calls to external sources can be routed through a policy-aware proxy. Exports for compliance, including full metadata backups or FOIA requests, are supported.

Security, Privacy, and Compliance Controls

Security and privacy are built into every layer. All communications are encrypted with TLS, while data at rest in S3 and Snowflake is encrypted with KMS-managed keys. Sensitive data handling is enforced by classification: masking or denial of PII/PHI is required before public publication, and restricted fields remain accessible only under approved policies. Licensing terms are mandatory, and external datasets include explicit storage of source terms. For restricted datasets, access requests may be gated by interstitials requiring terms acceptance. Retention is tracked at the dataset level, with scheduled reviews triggering archival or de-publication as necessary.

Operations & SLAs

Operations are governed by clear service levels. CKAN and PortalUS target 99.9% availability, while Snowflake is covered under its vendor SLA. External datasets are monitored for link health, and alerts are triggered when thresholds are exceeded. Metadata is backed up nightly, and Snowflake leverages point-in-time recovery and Time Travel for resilience. All configurations for CKAN extensions, metadata scheming, and authentication are managed as Infrastructure-as-Code, moving through dev, staging, and production pipelines with automated tests validating governance policies.

Roles & Responsibilities

The governance model is role-driven. Data Owners hold business authority, approving publication and retention. Data Stewards curate metadata, schemas, and quality, applying validation rules and

remediating issues. Compliance Reviewers confirm classification, licensing, and consent compliance. Platform Administrators manage CKAN, its extensions, logging pipelines, and integrations. Security Officers oversee identity, MFA enforcement, and incident response. This clear separation of responsibilities ensures that governance is shared but accountable.

Extensions & Components

The technical solution leverages CKAN's extensible architecture.

- Extensions such as scheming (for metadata models), harvesters (for catalog integration), xloader (for bulk dataset loads), archiver/qa (for availability and format validation), and dcat (for metadata interchange) are core to the deployment.
- Workflow and approval extensions enforce governance gates, while saml2/oidc enable Auth0 integration.
- Multilingual support ensures accessibility.
- Snowflake provides external stages, Snowpipe ingestion, dbt transformations, row-level security, and Time Travel for versioning.
- Policy-aware proxies secure external APIs and preview queries.
- Loguru integrates logging across the stack and consolidates events in Snowflake.
- Metabase embeds deliver governed analytics, including optional alerting via email or webhook tied to curated models.

Together, these components create a robust, modular, and scalable platform.

Content Management

In this architecture, CKAN serves as the authoritative system of record for all datasets. It manages metadata, provenance, and access control, while also providing REST and Action APIs that enable both search and detailed dataset retrieval. Complementing this, [Strapi.io](https://strapi.io) operates as the headless CMS responsible for all narrative and editorial content, including landing pages, news articles, blog posts, and dataset showcase stories. Unlike CKAN, Strapi is not intended to store raw data or dataset metadata but instead provides a content layer that enriches and contextualizes the datasets, while enabling additional features including announcements, blogs, and custom static pages.

The user experience is delivered through PortalUS, a modern React-based frontend that consumes APIs from both CKAN and Strapi. This allows the platform to present a seamless experience in which narrative content and dataset records appear integrated, even though they are managed by separate systems. To ensure secure access across both layers, Auth0 provides single sign-on (SSO), multi-factor authentication (MFA), and role-based access controls. For large-scale data hosting and curation, Snowflake serves as the data warehouse, with CKAN entries pointing to Snowflake-hosted objects.

Integration Patterns

Integration between CKAN and Strapi is achieved primarily through API-driven workflows. A lightweight mediation layer (either in the form of a Node.js service or a Strapi plugin) can be used to call CKAN's API on a scheduled basis and store lightweight dataset references (such as IDs, titles, and tags) inside Strapi. This ensures that CMS-managed content can include live connections to CKAN datasets without duplicating or conflicting with the canonical metadata.

From the editorial perspective, Strapi content models can include a "Dataset Reference" block. Editors can select or paste in a CKAN dataset ID, which then acts as a pointer to the authoritative record in CKAN. When these pages are rendered in PortalUS, the front end makes a live call to CKAN to retrieve the dataset details. This pattern ensures that Strapi content always displays the most current dataset metadata without requiring manual synchronization.

Sequence Flow

When a dataset landing page is created, the process begins with an editor authoring a new "Dataset Showcase" entry in Strapi. As part of the editorial workflow, the editor links this page to a specific dataset in CKAN by entering its dataset ID. Once the content is published, PortalUS requests the page from Strapi's API. The response includes the dataset reference, which PortalUS uses to trigger an additional API call to CKAN. CKAN then returns authoritative metadata, provenance details, and access information.

The result is a single rendered page where the narrative and contextual content are managed in Strapi, while the dataset details are delivered directly from CKAN. This approach keeps responsibilities

clearly separated: Strapi enriches and contextualizes content, while CKAN governs datasets and their metadata.

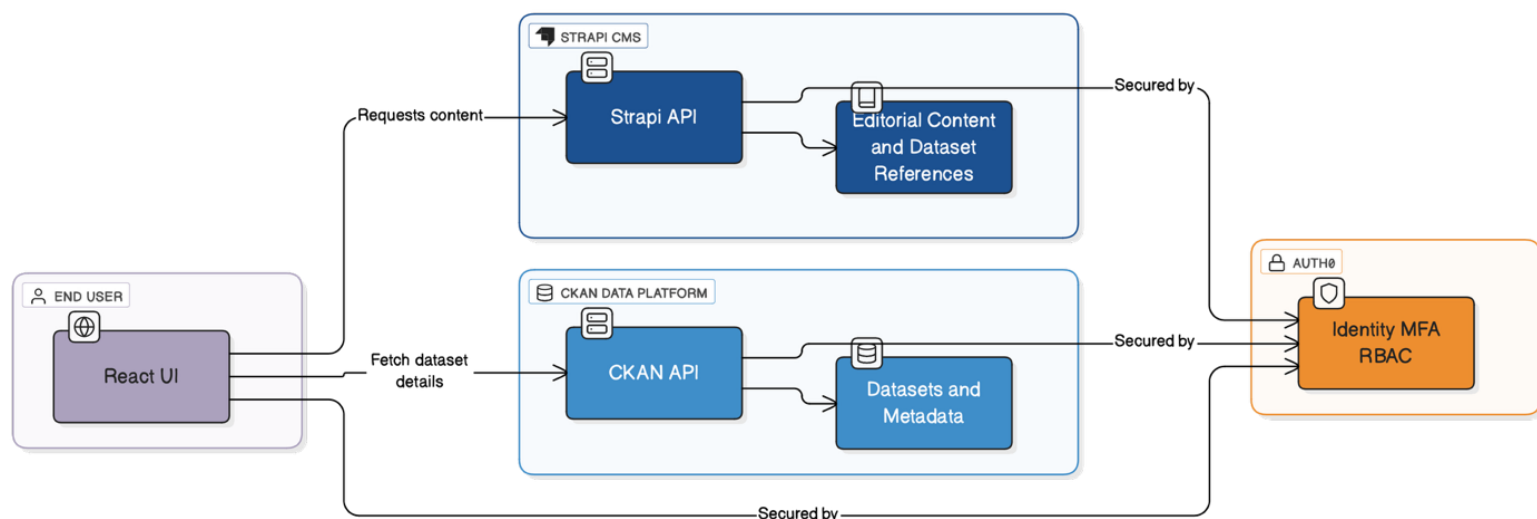


Figure 13 - CKAN + Strapi - Combined Content Delivery

Security and Governance

Auth0 sits across both CKAN and Strapi to ensure consistent security and identity management. Any call to the APIs is authenticated through a single identity provider, and sensitive datasets can be placed behind MFA or role-based approval workflows. This guarantees that both content and data benefit from the same access governance model, even though they are stored in separate systems.

This reference architecture establishes a clean separation of concerns: CKAN remains authoritative for data, Strapi governs content and storytelling, and PortalUS delivers a unified experience to end users. By leveraging APIs rather than monolithic integrations, the approach ensures extensibility, scalability, and future compatibility with other services such as analytics platforms, reporting tools, or additional CMS providers.

Translation & Internationalization

Translations.com's GlobalLink suite provides the backbone for our localization workflow. GlobalLink Project Director operates as the Translation Management System (TMS), orchestrating translation jobs, routing content through machine translation, machine-translation-plus-post-edit, or full human workflows, and applying translation memory and terminology for consistency. Project Director exposes robust APIs so the DCC can programmatically submit content, monitor progress, and retrieve translated assets.

For user interface text and other key-value strings, GlobalLink Strings (Applanga) offers a developer-friendly service with REST APIs and webhooks. It allows us to push and pull discrete strings, organize them into namespaces, and perform in-context review so that linguists can see how text appears within PortalUS before publication.

Underpinning both services are enterprise-grade language assets: a centralized termbase to enforce domain terminology and translation memory to reuse prior work, reduce cost, and accelerate turnaround. The wider connector ecosystem, and critically the availability of open APIs, means we can integrate cleanly even where there is no off-the-shelf connector for CKAN or Strapi.

Target Architecture

The DCC runs in AWS and introduces a small, stateless Localization Orchestrator that bridges DCC systems and GlobalLink. Strapi emits webhooks when content changes, CKAN exposes changes via API or plugin hooks, and PortalUS consumes string bundles directly. The Orchestrator converts rich text into XLIFF, submits translation jobs to Project Director, listens for callbacks, and writes the results back to Strapi and CKAN. For UI strings, it coordinates with Applanga to keep key sets synchronized and to react to webhook notifications.

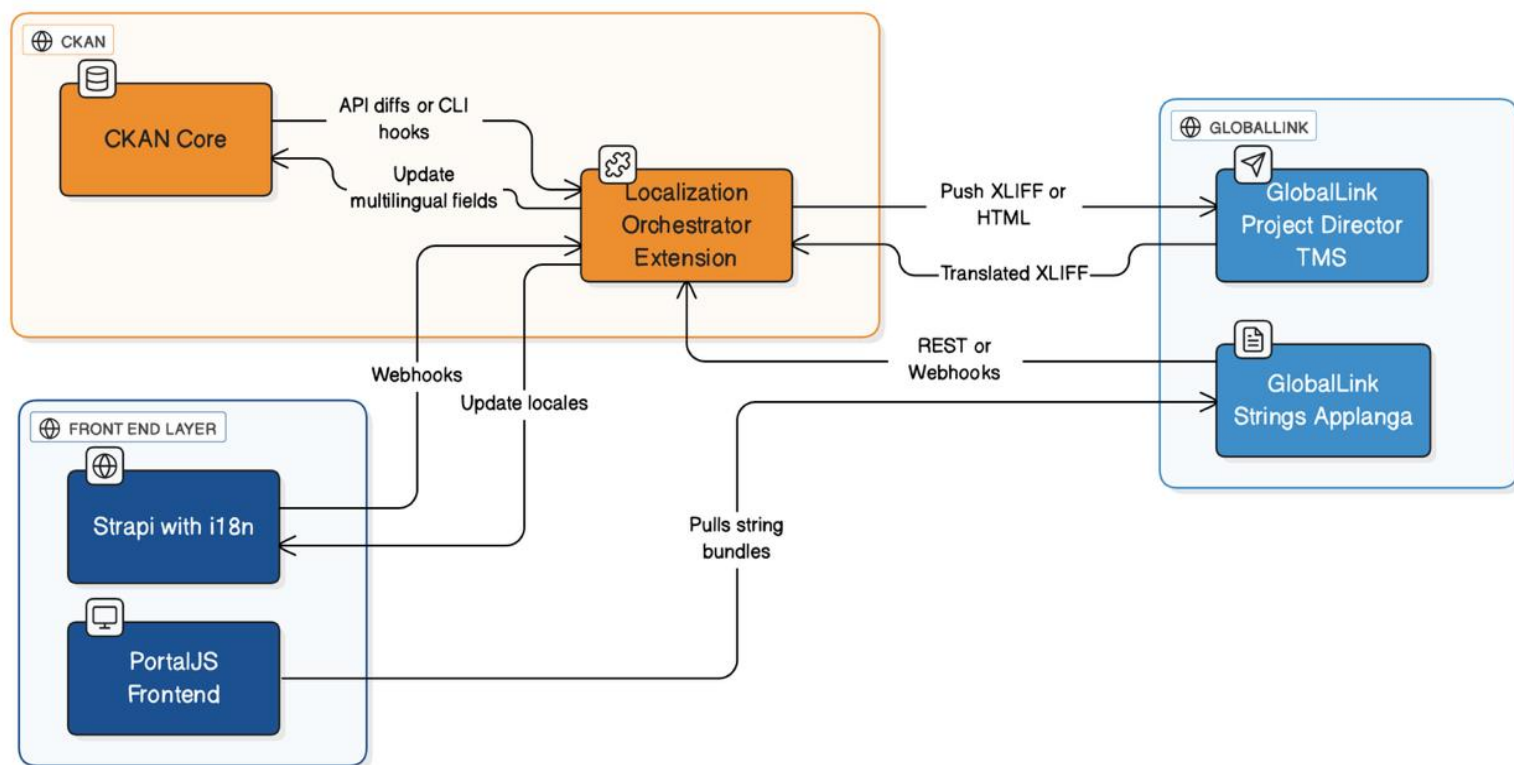


Figure 14 - Translation Reference Architecture

Components

Strapi, with its i18n capability enabled, remains the source of truth for all website content. Each localized value is stored natively, so authors work in familiar interfaces and see per-locale drafts and publication states. CKAN, extended with *ckanext-fluent* and *ckanext-scheming*, becomes the source of truth for multilingual catalog metadata such as dataset titles, descriptions, tag displays, and organization names.

PortalJS renders the public experience. It does not store translations; instead, it reads the correct string bundle from Applanga at build time or runtime, ensuring that UI chrome, filter labels, and microcopy are always in sync. The Localization Orchestrator, implemented in Node.js or Python and deployed on ECS Fargate or Lambda, manages extraction, submission, and re-ingest, while AWS services such as Secrets Manager, CloudWatch, and SQS/SNS provide security, observability, and resilience.

GlobalLink Project Director handles document-like content and structured metadata delivered as XLIFF or HTML fragments, whereas GlobalLink Strings (Applanga) manages key/value strings and notifies the Orchestrator about changes via webhooks. This separation lets each content type use the tool best suited to its lifecycle.

Data Flows

When editors create or update content in Strapi, a webhook notifies the Orchestrator, which retrieves the affected entry, transforms the changed fields into XLIFF, and submits a job to Project Director under the appropriate workflow. Once translation is complete, the Orchestrator retrieves the localized assets and updates the specific locale fields in Strapi, triggering a targeted CDN or cache invalidation so only the relevant pages are refreshed.

For CKAN, the Orchestrator detects changes through API polling or a lightweight extension hook, extracts translatable fields such as titles and descriptions, and submits them as XLIFF to Project Director. Completed translations are written back into CKAN's multilingual fields in a way that respects *ckanext-fluent*'s internal mapping so that search and browse experiences surface localized values correctly.

UI strings follow a complementary path. Developers or the Orchestrator push new or changed keys to Applanga. PortalUS then requests the appropriate bundle for each locale, and linguists can perform in-context review against the running UI. If keys are adjusted by reviewers, Applanga webhooks can trigger the Orchestrator to update any mirrored values in Strapi, keeping content and chrome aligned.

Integration Patterns

Strapi

Strapi content types that appear on the public site, including Announcements, Blog, Pages, and dataset narrative blocks, are localization-enabled so that each language variant lives alongside the source. Webhooks on create, update, and publish events send compact payloads that identify the entry, source locale, and changed fields. The Orchestrator fetches the canonical record, converts only the deltas into XLIFF 1.2/2.0 for safe round-trip, and maps each content type to a Project Director workflow that balances speed and quality. Low-risk content can auto-publish after machine translation plus light review, whereas policy pages and legal copy wait for human approval gates before going live.

CKAN

CKAN is extended with *ckanext-fluent* to provide multilingual fields and with *ckanext-scheming* to declare which fields are translatable. The Orchestrator either receives a signal from a simple CKAN plugin or performs scheduled diffs through the API to identify changed records. It assembles titles, notes, tags, and organization display names into XLIFF, submits them to Project Director, and writes the returned translations to the correct locale slots. CKAN's core UI translations continue to use the standard CKAN i18n packs; GlobalLink focuses on catalog content rather than application chrome.

PortalUS (UI Strings)

PortalUS manages UI strings via Applanga, organizing keys into stable namespaces that correspond to feature areas. At build time or on first request, the front-end loads the locale-appropriate bundle, and optional in-context review allows linguists to validate strings where they actually render. The Orchestrator listens for Applanga webhooks so that when reviewers adjust text, dependent content can be republished or caches pruned without broad invalidations.

Security & Compliance

All integrations communicate over TLS, and every webhook is signed and verified using HMAC with a timestamp and nonce to prevent replay. The Orchestrator assumes least-privilege IAM roles and loads credentials from AWS Secrets Manager with regular rotation. We avoid sending personally identifiable information (PII) in payloads; if business rules require it, we apply redaction or pseudonymization before submission. Every submission, callback, and publish event is logged with correlation IDs so that an auditor can reconstruct the lifecycle of a translation from source edit to localized publication.

DevOps & Observability

The Orchestrator is provisioned via infrastructure-as-code and deployed through CI/CD, with health checks and metrics published to CloudWatch. Transient failures, such as temporary API errors, are handled with exponential backoff and queued retries. Messages that cannot be processed are routed to a dead-letter queue for inspection. Feature flags allow us to roll out new locales gradually and to experiment with the proportion of content that uses machine translation versus human workflows.

Non-Functional Requirements

Turnaround times are tuned by content risk: human-reviewed items typically publish within twenty-four to seventy-two hours, while machine-translated content can propagate in under thirty minutes. The Orchestrator is stateless and horizontally scalable, with queueing to absorb spikes. Availability targets are set at 99.9% for the Orchestrator endpoints, and the platform degrades gracefully by serving source language content whenever a translation is missing.

Risks & Mitigations

Schema drift is the most common risk when content models evolve. We mitigate this with contract tests and XLIFF mapping unit tests in CI. Longer human translation cycles are addressed by using machine translation with post-editing for medium-risk content, parallel reviewer pools, and staggered locale rollouts. Webhook delivery failures are contained by idempotent handlers, exponential backoff, and dead-letter queues. To avoid inconsistent terminology across systems, we enforce termbase usage and mandatory translation memory application in Project Director.

Configuration Snapshot

Strapi is configured with i18n on all public-facing content types and with outgoing webhooks that send signed events on create, update, and publish. CKAN defines supported languages through *ckanext-fluent* and identifies translatable fields via *ckanext-scheming*, with an API key restricted to the Orchestrator's service account. The Orchestrator stores GlobalLink and Applanga credentials in Secrets Manager, verifies webhook signatures, enforces retry policies, and exposes health metrics. PortalUS declares Applanga namespaces, implements a predictable bundle-fetch strategy, and offers a locale switch that updates both UI strings and content without a full page reload where feasible.

AI Fairness and Bias Framework

Our team will deliver a comprehensive AI Fairness and Bias framework aligned with the Massachusetts AI Hub's mission to ensure that the Data Commons Collaborative (DCC) enables equitable, transparent, and trustworthy AI development. Our approach will integrate:

- **Human-Centric Design** – Safeguarding individual dignity, rights, and autonomy. AI systems will be designed to augment, not replace, human decision-making, with particular focus on public good applications.
- **Transparency & Accountability** – Embedding fairness dashboards, explainability tools, and auditable decision logs to ensure that researchers, policymakers, and the public can trust DCC-enabled AI outcomes.
- **Inclusive & Equitable Access** – Ensuring datasets and models reflect diverse populations. Community engagement mechanisms will be built into the governance workflow, including participatory audits and stakeholder feedback loops.
- **Sustainability & Innovation** – Leveraging open-source fairness tools and scalable monitoring pipelines to future-proof the DCC while supporting energy-efficient model development.

For the initial release of the DCC, we proposed the following solutions to support the reporting and governance of AI Fairness.

- **Detection & Auditing:** [AIF360](#) + [Fairlearn](#)
- **Mitigation:** AIF360 + Fairlearn
- **Explainability:** [SHAP](#) + LIME
- **Monitoring & Reporting:** [Evidently AI](#) + [WhyLogs](#)
- **Workflow Integration:** [MLFlow](#) + [Jupyter](#)

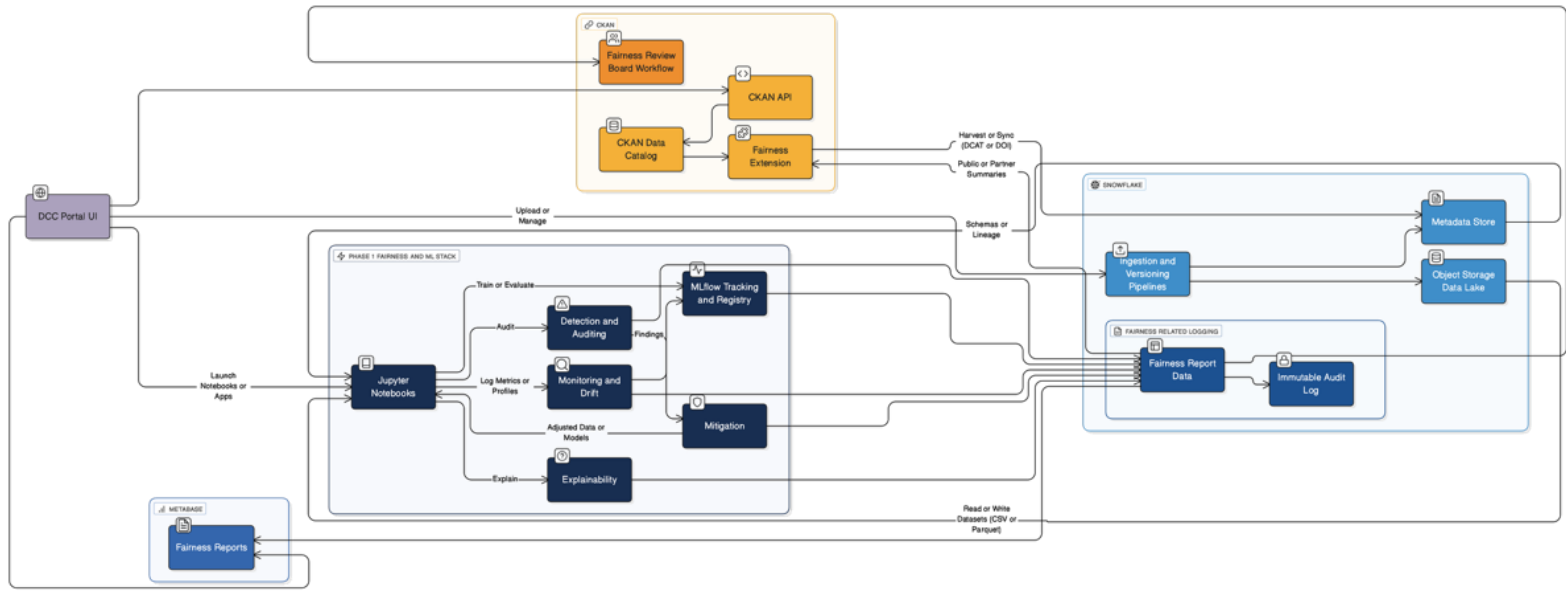


Figure 15 - AI Fairness Integration Architecture

Bias Detection & Dataset Auditing

The DCC will include robust bias detection and dataset auditing capabilities. Both automated and manual processes will be applied to identify disparities across protected attributes such as race, gender, age, and disability. Using statistical fairness metrics and visualizations, these audits will highlight underrepresented groups and risks of reinforcing inequities. Open-source tools AIF360 and Fairlearn will be integrated into the workflow to provide a comprehensive and transparent assessment of datasets.

Model Fairness Evaluation

Fairness evaluation will be deeply embedded within machine learning workflows. Researchers and developers will be able to assess model performance across different subpopulations directly within environments using Jupyter notebooks and MLFlow pipelines. Automated fairness reports will be generated, providing clear and actionable insights into how models behave for various groups, ensuring transparency in performance outcomes and promoting accountability in high-impact applications.

Bias Mitigation Tools

A full suite of bias mitigation approaches will be implemented across the data and model lifecycle. Pre-processing techniques such as dataset reweighting and re-sampling will improve representativeness of training data. In-processing methods, including fairness-aware algorithms and adversarial debiasing, will help train models that balance accuracy with equity. Post-processing techniques will allow adjustments to outcomes where disparities persist, with all tradeoffs carefully documented to ensure transparency for decision-makers.

Monitoring & Reporting

The DCC will offer continuous monitoring of fairness metrics through intuitive dashboards and

reporting. These dashboards will track fairness outcomes across projects and provide version-controlled histories of all audits and mitigation actions. Real-time alerts will notify users when performance or fairness drift occurs, ensuring that issues are addressed quickly. Tools such as Evidently AI and WhyLogs will be used to detect changes in data quality or fairness performance, maintaining accountability throughout the lifecycle of data and models.

Community Engagement

Recognizing that fairness extends beyond metrics, the DCC will incorporate mechanisms for meaningful community engagement. Impacted communities will be invited to participate in audits and provide feedback through accessible channels built into the platform. Documentation of social and historical context will accompany datasets and models, ensuring that lived experiences inform technical evaluations. Public-facing fairness summaries will be published to enhance transparency and build trust among users and stakeholders.

Legal & Regulatory Compliance

The fairness framework will align with state and federal anti-discrimination laws, as well as emerging regulations such as the Algorithmic Accountability Act and the EU AI Act. Governance structures will clearly define roles for fairness auditors, data stewards, and reviewers, with a designated Fairness Oversight Board responsible for high-risk use cases. Consent and opt-out mechanisms will be integrated into the data access process, ensuring compliance with both ethical and legal standards.

Education & Capacity Building

To ensure long-term sustainability, the DCC will provide extensive education and training resources on fairness and bias. Users will have access to online modules, workshops, and hands-on tutorials that explain key fairness principles and demonstrate practical application of fairness tools. A library of case studies and best practices will be maintained to support knowledge-sharing across the ecosystem. These efforts will create a community of practice in Massachusetts, positioning the state as a leader in responsible AI.

Synthetic Data Capability

The Synthetic Data capability will enable the DCC to generate privacy-preserving, high-utility datasets across multiple modalities while meeting strict governance, audit, and compliance requirements. The approach ensures that sensitive or partner data can be safely leveraged by creating synthetic equivalents that maintain statistical fidelity and analytic utility without exposing the original data. Synthetic outputs are clearly tagged, validated, and published in CKAN with complete provenance, meeting the requirements outlined in the DCC RFP.

Reference Architecture

Synthetic Data generation is fully integrated with the DCC architecture. Source data is first profiled and risk-scored, then passed into synthesis tools executing in containerized environments on AICR compute. Outputs undergo validation for utility, privacy, and fairness before being stored in S3/Snowflake and registered in CKAN. Auth0 enforces role-based access to synthetic datasets, while usage and quality metrics are captured in Metabase dashboards and all events logged through Loguru into Snowflake.

Tools by Modality

The platform will support multiple modalities:

- Tabular/Time-Series: SDV (CTGAN, TVAE, Copulas), with YData or Gretel for advanced privacy modes.
- Health: Synthea for realistic EHR/FHIR data.
- Autonomy/Robotics: CARLA for driving simulation; Virtual KITTI/UnrealCV for scenario expansion.
- Vision: Unity Perception and Sim4CV for computer vision training data.

Orchestrated Workflow

- 1) Register source dataset in CKAN with sensitivity and intended synthesis purpose.
- 2) Automated profiling and privacy risk scoring.
- 3) Generator selection based on modality and governance policies.
- 4) Containerized execution of synthesis jobs on AICR.
- 5) Validation of results for utility, privacy, and fairness.
- 6) Publication of synthetic dataset in CKAN with full lineage and reports.
- 7) Access enforcement through Auth0 and continuous logging via Loguru into Snowflake.

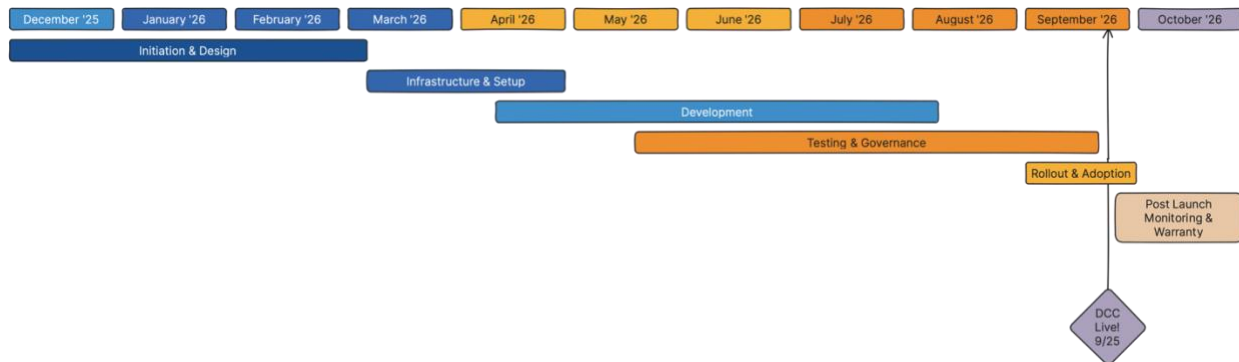
Governance & Compliance

Synthetic Data generation will be tightly coupled with governance processes. Datasets require approval before synthesis, and outputs undergo privacy and fairness validation. Synthetic datasets are typically published at a 'Registered' level but may be promoted after governance review. Audit artifacts and lineage records are retained to meet state, federal, and partner compliance requirements.

Commons Credits Integration

Synthetic Data creation will also tie into the Commons Credits framework. Contributors can earn credits for publishing high-quality synthetic datasets or validated generation recipes. Credits can be spent on GPU synthesis time, priority processing, or access to premium synthetic corpora. All activity is tracked in the Commons Credits ledger.

Phases, Scoped Activities, and Deliverables



Phase 1 – Initiation & Design

Strategy and Requirement Alignment

This phase validates user needs, market trends, and technology options through research and early prototyping of key use cases. By grounding the vision in real-world customer experience, it ensures the Data Commons is built around the highest-value opportunities, reducing risk and building early stakeholder support.

- Conduct stakeholder interviews and surveys (AI developers, researchers, government, startups).
- Benchmark comparable data commons (e.g., Australian Research Data Commons, CKAN, Dataverse).
- Pilot customer experience (CX) concepts with targeted use cases (e.g., health, transportation, climate).
- Develop user personas and journey maps.
- Draft strategic brief outlining vision, use cases, and early dataset priorities.
- Review and finalize Dataset Criteria for this initial phase of the engagement.

Milestones Addressed in this Phase

- Engagement Kickoff (team aligned, scope and goals confirmed)
- Dataset Curation Criteria Approved
- Initial Dataset List Confirmed

Activity Deliverables

- Program Kickoff Deck
- Data Commons Comparative Analysis
- Defined Dataset Curation Criteria
- Synthesized Stakeholder Feedback

- Strategic Brief

Detailed Requirement Gathering

Through stakeholder workshops and technical assessments, this phase translates strategy into clear functional, technical, and governance requirements. By aligning user expectations with compliance and FAIR data standards, it creates a strong blueprint that ensures feasibility, trust, and long-term usability.

- Facilitate design workshops to capture technical and governance requirements.
- Define dataset inclusion and curation criteria (public, synthetic, Massachusetts-specific).
- Establish metadata standards (DCAT, Schema.org, Dublin Core).
- Identify governance, security, and compliance needs (RBAC, MFA, encryption, HIPAA/state law).
- Document and validate detailed functional and user requirements specification.

Milestones Addressed in this Phase

- Detailed Architecture Defined (finalized technical design)

Activity Deliverables

- Architecture Document
- Finalized Requirements Documentation
- Development Backlog with Epics & User Stories

Program Planning

This phase defines the roadmap, resource plan, and delivery cycles while embedding risk management and governance checkpoints. By setting clear milestones and transparent resourcing, it enables predictable delivery, optimizes investment, and aligns stakeholders on a common vision for execution.

- Build program roadmap with milestones and releases.
- Estimate resource levels and effort by workstream.
- Develop iterative delivery plan (sprint cycles, feature releases).
- Define risk management strategies.
- Secure approval of program charter and work plan.

Milestones Addressed in this Phase

- Detailed Governance Defined (finalized governance policies)

Activity Deliverables

- Program Roadmap
- Industry Engagement Plan

- Steering Committee Meeting Cadence
- Deliverable Review Sessions Mapped

Phase 2 – Infrastructure & Setup

Environment and 3rd Party Service Setup

This phase establishes the technical foundation necessary to support development, testing, and future production operations of the Data Commons Collaborative. Core activities include provisioning secure cloud environments, configuring identity and access management, and enabling baseline monitoring and logging. In parallel, third-party services—such as metadata management platforms, synthetic data engines, and supporting SaaS integrations—are deployed and validated to ensure interoperability with the DCC architecture. Completing this phase ensures that the infrastructure and external tools are production-ready, compliant with governance requirements, and aligned to the scalability, security, and extensibility goals of the program.

Milestones Addressed in this Phase

- Platform infrastructure provisioned
- Identity and Access Management Setup with Auto
- Dev/test environments ready
- 3rd Party Core platform services deployed
- Foundations for security foundations, governance, and policy enforcement
- Validation and handoff for development

Activity Deliverables

- Provisioning of cloud infrastructure and identity management
- Stand up of development and test environments
- Security and Governance Implementation (initial RBAC, MFA, audit logging)
- Architecture Documentation (foundation services, identity mgmt.)

Experience Design

The Experience Design phase ensures that the Data Commons Collaborative delivers a clear, intuitive, and engaging user experience for its diverse audience of researchers, developers, and stakeholders. This phase focuses on translating functional requirements into a user-centered design framework that balances accessibility, usability, and visual appeal. Key activities include:

- **Information Architecture** – Defining the structure, navigation, and content hierarchy to support intuitive discovery and interaction.
- **UX Wireframing** – Creating user flows and low-fidelity prototypes to validate usability and align on functionality before development.

- **Visual Design** – Developing a cohesive visual language and interface style that reflects the DCC’s mission while maintaining consistency across all digital touchpoints.

Activity Deliverables

- Annotated UX Wireframes
 - Home
 - Search
 - Profile
 - Dataset Overview
 - Dataset Detail
 - Admin
 - Reporting
- Visual Designs for
 - 4 Page Templates
 - 12 Components

Phase 3 – Development

Agile Solution Development

The solution is built in iterations and reviews are scheduled every 3 weeks to demonstrate progress. This agile approach allows for rapid validation with users, course correction as needed, and continuous alignment to evolving needs, ensuring the solution is both innovative and practical.

We estimate 6 development Sprints for the full development scope. Each development team will work against the same sprint model and timeline so reviews will contain all the features developed in that Sprint by all teams.

- Implement dataset ingestion pipelines and metadata catalog.
- Integrate synthetic data generation tools (e.g., SDV, Synthea, CARLA).
- Build role-based access, Commons Credits ledger, and governance dashboards.
- Develop UX/UI and conduct usability testing.
- Perform sprint reviews with stakeholders.

Milestones Addressed in this Phase

- Data integration complete (ingestion pipelines, metadata catalog)
- Synthetic data engine operational (at least one tool producing test outputs)
- Website functional (browsing/search, dataset detail pages)

- Admin dashboard live
- Fairness and Bias Reporting Dashboard
- Security and Governance Implementation
- Training Materials Developed

Activity Deliverables

- Initial Data Catalog and Metadata System
- Synthetic Data Tooling and Sample Outputs
- Sample Commons Credits Ledger Output
- Fairness and Bias Reporting Dashboard
- Security and Governance Implementation
- Training Materials
 - Administrator Guide
 - Administration Training Presentation
 - 2 Administrator Training Session
 - Documented Training Guide (for Train the Trainer support)
 - End User Training Guide
 - 2 End User Training Sessions

Phase 4 – Testing & Governance

QA & Performance Testing

Rigorous QA and performance testing ensures the Data Commons is stable, secure, and scalable before launch, reducing the risk of failures in production. This builds user trust, safeguards sensitive data, and provides confidence that the platform can handle real-world demand.

The rollout plan includes end-to-end QA testing (functional, performance, and security) prior to launch. Training sessions will be conducted for administrators and end-users, accompanied by user guides and video tutorials. Pilot testing with selected researchers and civic tech partners will provide feedback loops to refine features before public launch. Launch materials will include a communications toolkit to promote adoption across MassTech's AI ecosystem.

Milestones Addressed in this Phase

- Platform QA complete (functional, performance, security testing)
- Governance framework operational (role-based access, usage agreements)

Activity Deliverables

- Pilot user feedback

Security & Penetration Testing

Engaging a third-party for security and penetration testing is a critical best practice to ensure the integrity and security of any system, including the DCC. Independent testing provides an unbiased, expert evaluation of the system's vulnerabilities, offering a fresh perspective beyond internal teams. It helps identify potential security risks, weaknesses in the application, and flaws in the infrastructure before they can be exploited. This proactive approach not only strengthens the DCC's security posture but also aligns with industry standards, mitigates risk, and ensures the system is resilient against external threats. By engaging third-party experts, we ensure a comprehensive and thorough security review that is essential for maintaining user trust and safeguarding sensitive data.

We will partner with the 3rd party vendor to review and address any feedback or identified areas of concern.

Activity Deliverables

- Resolution of any identified security consideration

User Acceptance Testing (UAT) with Stakeholders & Pilot Users

The User Acceptance Testing (UAT) phase occurs after all Quality Assurance (QA) activities are complete. Unlike QA, which validates technical functionality, UAT is designed as a pre-launch validation step that focuses on the experience of stakeholders and their readiness to assume operational responsibilities.

During UAT, the Steering Committee and designated stakeholders will interact with the DCC in its near-final form. The purpose is not to execute test scripts, but to provide a controlled opportunity for users to validate that the platform operates as expected for their roles and responsibilities, and to ensure that no user experience issues remain before public launch.

Objectives of UAT

- **Stakeholder Validation** - Confirm that the platform supports stakeholders in carrying out their expected roles, such as data submission, governance approvals, dataset discovery, or credit usage.
- **User Comfort and Readiness** - Ensure stakeholders are confident in using the platform and fully understand how to perform their key responsibilities.
- **Final User Experience Review** - Identify and resolve any outstanding usability or interaction concerns that could hinder adoption.
- **Launch Readiness Assurance** - Provide Steering Committee members and other key stakeholders with assurance that the DCC is prepared for public launch.

Process

- UAT is conducted in a production-ready environment configured with representative datasets, governance workflows, and Commons Credits functionality.
- Stakeholders will carry out typical day-to-day interactions (e.g., browsing datasets, validating governance workflows, reviewing dashboards) to ensure the platform is intuitive and aligned with expectations.
- Any issues or refinements identified will be documented and prioritized for resolution before launch.
- UAT will conclude with a formal review and sign-off by the Steering Committee, confirming that all stakeholders are confident in moving to production.

Activity Deliverables

- UAT summary report (capturing feedback, refinements, and resolutions)
- Steering Committee sign-off to begin launch steps
- Completed Training sessions

NOTE: Since our QA phase focuses on the functionality and quality of the solution, we have included a slightly condensed UAT window to support a Q3 launch date. If preferred, we can extend this UAT window to allow for more stakeholder engagement during this timeframe.

Phase 5 – Rollout & Adoption

Product Launch

By validating the solution with real users, this phase ensures the Data Commons is secure, accessible, and trusted, driving adoption across the Commonwealth and creating a foundation for sustainable growth and innovation. A soft-launch or Beta test is recommended to validate the product and ensure confidence for expanded launch.

Milestones Addressed in this Phase

- Run beta pilots with real users across academia, startups, and state agencies.
- Deliver user training, onboarding, and support documentation.
- Launch platform with public engagement strategy.

Activity Deliverables

- User Onboarding & Training Delivered
- Public Launch

Post Launch Monitoring & Warranty

To ensure stability and user confidence following the launch of the Data Commons Collaborative, we will provide a 30-calendar-day warranty period. During this window, any bugs or defects reported will be fully covered under the warranty, regardless of when the resolution occurs. This means that if an

issue is identified within the warranty period, our team will address and resolve it at no additional cost, even if the resolution extends beyond the 30-day timeframe.

The warranty applies strictly to defects or errors that prevent the solution from functioning as designed and delivered. It does not cover requests for new features, enhancements, or changes in scope, which would be managed separately through the project's change request process.

Recommended Services & Platforms

Solution Analysis & Justification

To provide transparency and confidence in the recommended architecture, we have included this section to highlight the reasons why each solution was recommended.

Comprehensive Knowledge Archive Network (CKAN)

We will use CKAN (Comprehensive Knowledge Archive Network) as the Foundation for the Massachusetts AI Hub Data Commons Collaborative (DCC). CKAN the world's leading open-source data management and data portal platform. It was designed to help organizations publish, manage, and share datasets, providing both a user-friendly web interface and a powerful backend for metadata management, governance, and programmatic access. CKAN has become the global standard for data commons and open data initiatives, making it an ideal foundation for the Massachusetts AI Hub Data Commons Collaborative (DCC).

We have included CKAN as the foundation for the DCC due to its alignment with key requirements:

1. **Web Experience and Portal Functionality**

CKAN delivers a rich, customizable web experience that supports:

- A searchable and filterable dataset directory
- Dataset detail pages with metadata, documentation links, and related resources
- Community dataset submissions and administrative dashboards
- Responsive design with accessibility compliance

These features directly align with the DCC's requirements for a website, dataset directory, dataset detail pages, admin dashboard, and public access portal.

2. **Metadata Management and Discovery**

CKAN natively supports FAIR data principles and leading metadata standards, including DCAT, Dublin Core, and Schema.org. Its robust cataloging and faceted search capabilities allow users to easily discover and reuse datasets. This maps precisely to the DCC's requirement for metadata management, dataset publishing, discovery, versioning, and governance.

3. **Security, Governance, and Access Control**

CKAN includes built-in role-based access control (RBAC), dataset tiering, and audit logs. It supports differentiated access levels (public, registered, restricted), which are critical for ensuring secure, ethical, and compliant data use as required in the RFP.

4. **APIs and Integration Capabilities**

CKAN provides a comprehensive RESTful API that enables developers, researchers, and

government partners to integrate directly with the platform. This ensures that datasets can be accessed programmatically for research, innovation, and application development.

5. **Extensibility for Synthetic Data and Commons Credits**

CKAN provides both a modular plugin architecture for extension and a robust event framework, referred to as Signals. The Signals framework provides native events for common actions such as dataset creation and is easily extensible to provide events for custom actions such as redeeming credits. With its modular plugin architecture, CKAN can be extended to:

- Integrate with synthetic data generation frameworks (e.g., Synthea, SDV, CARLA)
- Support usage analytics and monitoring
- Enable implementation of the Commons Credits ledger system to track contributions, incentivize participation, and ensure sustainable engagement.

Additionally, CKAN is the most widely adopted open-source data portal solution globally, with deployments by:

- **U.S. Federal Government** – [Data.gov](https://data.gov), the nation’s open data portal
- **European Union** – the [EU Open Data Portal](https://data.europa.eu).
- **United Kingdom** – data.gov.uk, one of the earliest and most extensive national open data portals.
- **State and Local Governments** – dozens of portals across U.S. states and major cities such as [Boston](https://data.boston.gov)

This widespread use demonstrates CKAN’s scalability, reliability, and suitability for public sector data governance and open access initiatives. CKAN provides the Massachusetts AI Hub Data Commons Collaborative with a proven, extensible platform that delivers the web experience, portal functionality, metadata cataloging, governance features, and security controls required in the RFP. Its global adoption by state, local, and national governments ensures a trusted, future-proof foundation for the DCC, enabling Massachusetts to lead in ethical, secure, and innovative AI-driven research and development.

Data Management Platform Comparison

For reference, we did consider the other solutions you listed in the RFP. While each has their strengths, none of them aligned to the core DCC requirements as strongly as CKAN.

Platform	License	Underlying Tech Stack	API Support	Customization & Extension Approach	Known gov use	Strengths	Weaknesses
Amundsen	Open Source (LF)	Python/Flask, Neo4j, Elastic, React	REST & GraphQL	Databuilder, UI tweaks	Few/no gov	• Easy to deploy • Strong UI • Lineage	• Smaller community • Less governance • No recent updates to the core
CKAN	Open Source	Python, PostgreSQL, Solr/Elastic	REST APIs (GraphQL available via extension library)	Plugins, theming, custom fields	Yes (Data.gov, EU, UK, AU)	• Government-grade open data • Extensible	• Limited lineage/analytics
DataHub	Open Source (LF)	JVM microservices, Kafka, React	GraphQL, REST, Python SDK	Ingestion pipelines, extensible metadata	Few/no gov	• Modern metadata • Lineage • Ecosystem integrations	• Complex infrastructure • Not portal-focused
Dataverse	Open Source (Harvard)	Java EE, PostgreSQL, Solr	REST APIs	Metadata schemas, storage integrations	Yes (Harvard, Dutch, Canadian)	• Preservation • DOI integration • Research community	• Heavy stack • Research-centric

Figure 16 - Data Management Platform Comparison

Auth0

The RFP calls for a robust role-based access and identity management system, multi-factor authentication (MFA), and secure governance capabilities for the DCC.

Auth0, an industry-leading identity platform provided by Okta, addresses these requirements directly:

- **Role-Based Access Control (RBAC):** Auth0 enforces fine-grained access policies aligned to DCC user roles (Public Viewer, Researcher, Curator, Administrator).
- **Single Sign-On (SSO):** Auth0 supports SAML, OIDC, and OAuth2, enabling integration with university, government, and enterprise identity providers for federated access.
- **Multi-Factor Authentication (MFA):** Configurable MFA (SMS, TOTP, push) ensures additional security when accessing restricted datasets.
- **Audit Logging & Compliance:** Auth0 provides detailed authentication and authorization logs, retained to meet the DCC's 3+ year audit requirement.
- **Governance and Security Alignment:** Auth0 ensures adherence to state and federal data security standards, supporting ethical and transparent governance.

This ensures that authentication, authorization, and governance are consistently enforced across all DCC components. Adopting Auth0 offers the Massachusetts AI Hub a secure, scalable, and sustainable solution for identity management:

1. **Security and Compliance by Design** – Auth0 is certified against SOC 2, HIPAA, FedRAMP, and GDPR, reducing compliance risks and costs.
2. **Rapid Implementation, Lower Development Risk** – Eliminates the need for custom identity systems, accelerating launch while minimizing engineering overhead.
3. **Scalability for a Growing Ecosystem** – Supports millions of users, ensuring seamless growth from Massachusetts pilot users to a global AI research audience.
4. **Flexibility for Federated and Tiered Access** – Natively supports federated logins and tiered dataset permissions, aligning with DCC governance needs.
5. **Long-Term Extensibility** – Robust APIs and SDKs ensure integration with future DCC modules (e.g., blockchain audit trails, consent management).
6. **Proven Reliability** – Auth0's widespread use in government, academia, and industry reduces risk and builds trust among stakeholders.

By leveraging Auth0, the DCC secures a trusted, future-ready identity layer, ensuring data access is governed ethically, securely, and at scale.

Okta vs. Auth0 by Okta

As people tend to be more familiar with Okta than Auth0, we have included the comparison below to highlight the differences and key use cases for each solution.

Dimension	Okta	Auth0
Primary focus	Workforce Identity & Access Management (internal employees, contractors, partners)	Customer Identity & Access Management (end-users of apps, SaaS, consumer platforms)
Target users	IT teams, CIOs, HR system admins	Developers, product teams, SaaS builders
Strengths	- Enterprise-grade SSO, MFA, lifecycle management - Okta Integration Network (thousands of pre-built connectors) - Strong compliance (FedRAMP, HIPAA, etc.)	- Developer-first approach (SDKs, APIs, CLI tools) - High customization for login flows, rules, and hooks - Flexible CI/CD integration for app developers
Customization	Config-driven, with strong admin UI and policies	Code-driven, extensible via rules, actions, and custom APIs
Integration style	Pre-built SaaS connectors, HR-driven provisioning	SDKs & APIs for embedding identity directly into sites and apps
Deployment model	SaaS-first, optimized for IT-managed environments	Flexible multi-cloud support, built for modern SaaS and global apps
Common use cases	- Single sign-on for employees - HR-driven user provisioning - MFA for internal systems - Partner/vendor access	- User registration and social login - API security with JWT/OAuth - Multi-tenant SaaS authentication - Consumer-facing app identity

Figure 17 - Auth0 vs. Okta

bridge.xyz

We proposed to implement the Commons Credits system using Bridge.xyz, a stablecoin issuance and management platform backed by Stripe. Bridge.xyz provides the security, compliance, transparency, and future-proofing required to support the Massachusetts AI Hub Data Commons Collaborative (DCC).

Security & Risk Management

Bridge leverages an immutable blockchain ledger to ensure that all Commons Credit transactions are tamper-proof, auditable, and transparent. In addition to blockchain's intrinsic security benefits, Bridge provides advanced controls such as account whitelisting and transaction monitoring. These features minimize exposure to unauthorized activity and reduce financial and reputational risk.

Secure Wallet Storage

Bridge provides a regulated custody solution for stablecoins, offering MassTech a secure and compliant method for managing Commons Credit balances. Wallet security is backed by Bridge's

technical infrastructure and financial licenses, ensuring that Commons Credits are protected from both operational risks and external threats.

Compliance

Bridge is a registered Money Services Business (MSB) and holds Money Transmitter Licenses (MTLs) across the United States. The platform provides:

- Know Your Customer (KYC) compliance to verify identity and prevent financial crime
- Monthly summaries of transaction activity, balances, and earnings for program oversight
- Quarterly attestations of tokens and reserves by an independent third party, ensuring that all issued Commons Credits are fully backed and accurately reported.

Established Provider

Bridge.xyz is operated by Stripe, a global leader in payments and financial infrastructure. Stripe's extensive expertise, strong reputation, and demonstrated operational rigor provide confidence in the stability and reliability of the Commons Credits program.

Transparency

Bridge provides a real-time reporting API, enabling MassTech to monitor:

- The total supply of Commons Credits in circulation
- The underlying reserve assets, which are fully backed by U.S. dollars
- The available inventory for redemption and swap

This transparency ensures that program stakeholders, auditors, and end-users have continuous visibility into the health of the system.

Future-Proofing: Third-Party Payments

By issuing Commons Credits as a USD-backed stablecoin, MassTech positions the system for broader ecosystem integration. Commons Credits could be applied toward third-party services (e.g., cloud or AI tool subscriptions such as Hugging Face Pro). Even if third parties do not directly accept the credit token, balances can always be redeemed for U.S. dollars, ensuring flexibility and utility.

Future-Proofing: Multi-Currency Support

Bridge's platform supports multi-currency stablecoins, allowing Commons Credits to eventually be purchased in other local fiat currencies. This capability creates a pathway for international expansion and broader participation in the DCC.

Revenue Generation: Earned Interest

Balances held with Bridge are invested in cash and short-term U.S. Treasuries. Any earnings from these reserves are shared with MassTech, providing a sustainable revenue stream that can be reinvested into the DCC to support future growth and innovation.

Metabase

For the Data Commons Collaborative (DCC), we recommend Metabase as the reporting and analytics solution. Metabase is a modern, extensible business intelligence platform that can natively connect to multiple data sources, ensuring flexibility for future growth and integration needs. The majority of reporting will be driven from Snowflake, and Metabase offers robust, optimized connectors for Snowflake that ensure secure, performant access to curated datasets.

We specifically recommend the SaaS Pro tier or higher of Metabase rather than the open-source edition. While the open-source version provides core reporting capabilities, it only supports embedding charts through IFRAMES. In contrast, the SaaS Pro edition provides modern React-based embedding options, which aligns with the DCC's requirements for a seamless, modular, and extensible user experience within the portal. This enables more secure and flexible integration of charts and dashboards, ensuring that analytics feel native within the Data Commons environment.

In addition, the SaaS editions of Metabase support alerts and subscriptions. Users can configure automated notifications when metrics change or cross thresholds, with delivery options via email, Slack, Microsoft Teams, or webhooks. This capability strengthens governance and transparency by enabling administrators and stakeholders to be proactively notified about critical dataset usage, Commons Credits balances, or key performance indicators. It also supports adoption by keeping researchers, startups, and government agencies engaged with real-time insights.

Metabase directly supports the DCC requirements outlined in the RFP:

- **API-Based Data Access & Usage Analytics** – Metabase provides APIs to programmatically query, visualize, and embed results, supporting the DCC's requirement for API-based access and extensibility.
- **Metadata & FAIR Data Principles** – Metabase integrates seamlessly with Snowflake's metadata and versioned datasets, enabling curated, transparent reporting that aligns with FAIR principles.
- **Security and Governance** – With the SaaS edition, Metabase supports enterprise-grade security, role-based access control, audit logging, and alerts/notifications, aligning with DCC governance requirements (RBAC, MFA, and audit trails).
- **Extensibility for Future Needs** – The ability to connect to multiple back-end data sources ensures the DCC can expand reporting as new data platforms or synthetic data pipelines are introduced, meeting the requirement for a future-proof architecture.
- **User Experience & Adoption** – Embedded charts and proactive alerts improve accessibility for end users (researchers, startups, government, and academic users) while reducing the burden of manual monitoring.

In summary, selecting Metabase SaaS Pro ensures that the DCC will deliver a modern, secure, extensible, and proactive analytics capability that maps directly to the RFP's technical architecture, governance, and user adoption goals.

Snowflake

The Massachusetts AI Hub Data Commons Collaborative (DCC) requires a secure, scalable, and widely adopted platform for hosting and managing data. Our recommendation is Snowflake, the state's existing enterprise data warehouse of choice. Snowflake's cloud-native architecture, strong security posture, and ease of integration make it uniquely aligned with the Commonwealth's requirements. Its widespread use across Massachusetts agencies ensures operational familiarity, streamlined procurement, and reduced onboarding time.

While Databricks was evaluated as a viable option—particularly given its strengths in advanced analytics and AI workflows—Snowflake offers a superior fit for the DCC's initial needs around curated dataset hosting, governance, and extensibility. Snowflake's strengths in structured data storage, compliance frameworks, and user scalability directly support the DCC's mission to centralize and democratize access to trusted data assets.

Role of Snowflake in the DCC Architecture

Snowflake will act as the foundational data warehousing layer within the overall architecture:

- **Dataset Hosting:** CKAN will serve as the data catalog and management interface, while Snowflake provides the underlying storage and compute capabilities for hosted datasets. This integration ensures that users benefit from a familiar CKAN interface while leveraging Snowflake's scalable query engine for discovery, exploration, and programmatic access.
- **Logging and Auditability:** Snowflake will serve as the central repository for all system logs. CKAN will persist operational and dataset-level logging into Snowflake, ensuring transparency and traceability of activity across the platform. Additionally, API-level logs from interactions with external services (e.g., Auth0 for authentication and consent management, bridge.xyz for Commons Credit transactions) will also be captured and stored in Snowflake. This provides a unified audit trail that supports compliance with state and federal security requirements.
- **Integration Enablement:** Snowflake will act as the bridge to downstream analytics, visualization, and monitoring frameworks. Reporting tools such as Metabase will connect to Snowflake to surface insights on dataset usage, governance, synthetic data generation, and credit transactions.

Benefits of Snowflake Adoption

Adopting Snowflake as the primary data warehouse provides several key benefits to the DCC:

1. **Alignment with State Standards** – Snowflake is already in active use across Massachusetts, creating immediate alignment with state IT strategy and procurement practices.

2. **Security and Compliance** – Built-in support for encryption, RBAC, MFA, and audit logging ensures compliance with state and federal privacy regulations.
3. **Scalability and Performance** – Elastic compute and storage provide flexibility to scale from pilot usage to statewide adoption without re-architecture.
4. **Unified Logging Repository** – Centralizing logs from CKAN and third-party services enhances governance, monitoring, and audit readiness.
5. **Future Extensibility** – Snowflake’s ecosystem of connectors and integrations positions the DCC for long-term evolution, including advanced AI/ML workloads that can complement synthetic data generation and fairness auditing.

Strapi.io

Strapi is a mature, open-source (MIT) headless CMS that cleanly separates content from presentation, which fits our PortalUS/React front end and CKAN-powered data catalog. We recommend Strapi Cloud to minimize ops overhead and keep costs predictable; current plans range from \$15 (Essential) to \$375 (Scale) per month, with a Free tier for non-prod uses. Strapi natively exposes REST and GraphQL APIs, supports webhooks, and provides a flexible content modeling system that lets us craft rich, multilingual pages (announcements, blogs, landing pages, and dataset-enhancement content) beyond what CKAN offers in its metadata UI.

Benefits of using Strapi.io

1. **Open-source preference & extensibility** - Strapi is open-source (MIT) and designed for extensibility (plugins/middleware), aligning to the DCC’s call for open-source technologies that promote transparency and extension.
2. **Website experience (homepage & directory)** - Requirements include highlighted collections, quick search/filter, recent updates, and a searchable dataset directory with cards and filters. These can be modeled as Strapi content-types for homepage modules, collection hubs, and dataset cards.
3. **Dataset detail pages** - DCC requires full metadata, related datasets, and external documentation links. Strapi will drive the experience wrapper—rich narrative blocks, callouts, images, video embeds—while CKAN remains the dataset system of record.
4. **Admin dashboard needs** - Add/edit content, manage user submissions, link health, and role-based access. Strapi’s admin, RBAC, and plugin model address these.
5. **Hosted-data UX enhancements** - Requirements include metadata standards, API access, version/lineage, previews, filters, and access-request flows. Strapi handles the site UX and

content (filters, request-access pages, explainer content), while CKAN and Snowflake handle data, lineage, and previews.

6. **Accessibility, performance, security, responsiveness, and extensibility** - Strapi's headless model lets us enforce accessibility via React components, secure APIs, and scale via Cloud plans; the architecture is modular for future APIs/data sources.
7. **Security, governance, SSO, and audit** - The DCC asks for RBAC, MFA/SSO, logging/audit, and compliance. Strapi provides RBAC in the admin, SSO (Enterprise or add-on), and Audit Logs (Enterprise); Auth0 complements these at the site edge.
8. **Internationalization** – Strapi.io supports the i18n for internationalizing and translating the experience.

Google Analytics

Google Analytics directly supports the Data Commons Collaborative (DCC) requirements by enabling measurement of user adoption, engagement, and dataset usage in line with MassTech's goals. The DCC requires strong usage analytics to understand how datasets, synthetic data tools, and governance features are accessed and applied.

Google Analytics provides event tracking, user journey mapping, and traffic insights that allow administrators to:

- Monitor dataset discovery, downloads, and API usage to validate adoption and FAIR data principles
- Assess community participation (e.g., submissions, feedback) and evaluate the effectiveness of governance policies
- Measure engagement across user groups such as researchers, startups, government agencies, and nonprofits
- Track performance, accessibility, and responsiveness of the web platform to meet operational requirements
- Provide reporting dashboards that align with transparency and audit needs while informing future enhancements

By offering scalable, cost-effective, and widely adopted analytics capabilities, Google Analytics ensures that the DCC can demonstrate value, optimize user experience, and guide data-driven improvements that align with its core mission of fostering ethical, secure, and innovative AI research.

Translations.com

We recommend Translations.com’s GlobalLink platform to meet the DCC’s translation and internationalization goals. GlobalLink combines a robust Translation Management System (Project Director) for documents and structured content with a developer-first strings service (Applanga) for UI text. Together, they deliver the automation, quality controls, and security DCC requires—while integrating cleanly with Strapi (site content), CKAN (catalog metadata), and PortaUS (front-end UI).

The following capabilities align to the requirements and long-term vision for the DCC and are the key pillars of our recommendation.

1. **One platform, two fit-for-purpose tracks** - DCC needs translation for (a) web content in Strapi and (b) catalog metadata in CKAN, plus (c) UI strings in PortaUS. GlobalLink covers documents/structured content via Project Director (TMS) and covers UI keys via GlobalLink Strings (Applanga)—both with mature APIs, so we can automate end-to-end without manual export/import.
2. **Developer-friendly automation for Strapi & PortaUS** - Applanga offers a REST API to push/pull string bundles and manage keys, and it supports webhooks so our orchestrator is notified the moment translations change—perfect for keeping PortaUS and any mirrored labels in Strapi in sync. Strapi already supports localization at the model/API level, so updates can be written back by locale.
3. **Clean path to multilingual CKAN metadata** - CKAN’s *ckanext-fluent* extension stores multilingual values for titles, notes, org names and tags; the DCC can extract source text to XLIFF, submit to Project Director, then write localized values back into CKAN’s multilingual fields—exactly what fluent is built to support. CKAN’s own multilingual guidance also documents how translated terms surface in search/browse.
4. **Quality, consistency, and speed knobs** - GlobalLink gives us translation memory and terminology to keep dataset names, tags, and site copy consistent across locales, and it lets us choose per-content workflows: MT, MT+post-edit, or human-only, depending on risk/importance. That directly supports DCC’s “right-sized” SLAs for policy pages vs. blogs vs. metadata.
5. **Scales with governance and integrations** - GlobalLink advertises 100+ prebuilt integrations and “robust APIs,” which reduces long-term integration risk and vendor lock-in. Even where we build a thin connector for CKAN/Strapi, we’re building on stable, documented APIs (Project Director, Applanga).

Alignment to DCC Requirements

Requirement: Multilingual site content (Strapi) with minimal friction.

Alignment: Use Strapi’s native i18n storage per locale; trigger a webhook to the DCC Localization

Orchestrator on create/update; submit XLIFF to Project Director; write back localized fields and republish. Strapi's i18n and GlobalLink's APIs make this a straightforward, event-driven loop.

Requirement: Localized catalog metadata (CKAN) that appears in search/browse

Alignment: Enable *ckanext-fluent* and declare translatable fields with *ckanext-scheming*; orchestrator extracts source text, sends to Project Director, and patches localized values back. CKAN's multilingual docs confirm localized names/descriptions surface when available.

Requirement: Localized UI chrome for PortalUS with fast iteration

Alignment: Manage keys/namespaces in Applanga; PortalUS loads bundles per locale; webhooks notify us on translation updates for instant cache/CDN refresh and optional in-context review.

Requirement: Quality controls, reuse, and cost/time optimization

Alignment: Apply translation memory and termbase across Strapi copy and CKAN metadata; select MT/MT+PE/human workflows per content type and locale to meet SLAs and budget.

Requirement: Open, API-driven integration and future-proofing

Alignment: GlobalLink offers documented APIs (including libraries for Project Director) and a broad connector ecosystem; Applanga's REST+webhooks model is simple to maintain and swap if the DCC's front-end stack evolves.

Translations.com (GlobalLink) meets the DCC where it operates: API-first, standards-aware (XLIFF), with automation for Strapi and first-class support for UI strings (Applanga) and CKAN metadata (via fluent). It gives us the quality levers (TM/termbase + workflow choices) to balance speed, cost, and accuracy—while keeping integration risk low thanks to mature, well-documented interfaces.

Team & Key Personnel

We are pleased to provide the initial set of profiles for team members who will be involved in our work together. These profiles represent leaders who will be involved directly and through the teams they lead along with key individual contributors.

Chitra Sundaram - Practice Director, Data Management

Professional Summary

Chitra Sundaram brings together business vision and technical expertise to lead enterprise data transformation. She has directed large-scale initiatives in data strategy, governance, and analytics across industries including retail, life sciences, and utilities. At Cleartelligence, Chitra partners with clients to shape data and AI roadmaps and leads teams to deliver platforms and solutions that drive measurable business value.



Chitra also leads the Data & AI Track for Sustainable Architectures, a non-profit who focuses on embedding sustainable practices into digital ecosystems helps minimize carbon footprints, promote responsible resource use, and encourage innovations that prioritize long-term well-being over short-term gains. She presented at the Sustainable Architectures Summit in London in June 2025 and has published articles in CIO.com, including [Why CDOs should prioritize ESG in data strategy](#) and [Beyond the hype: Do you really need an LLM for your data?](#)

Core Competencies

- Snowflake architecture
- Databricks architecture
- dbt
- Data ingestion frameworks & tools
- CICD with GitHub

Technical Skills & Platform Experience

- AWS, Azure
- Data lake architecture
- Postgres, DB2, Oracle, SQL Server
- Snowflake
- Databricks
- Informatica Data Quality, Replication
- Alteryx

John Fico – Practice Director, Software Engineering

Professional Summary

John Fico is a seasoned technical leader and consultant with over 25 years of experience architecting, developing, and delivering enterprise-scale software and data solutions. John specializes in full-stack application development, enterprise data platforms, and solution architecture, with deep expertise across cloud environments (AWS, Azure), databases, and modern development frameworks. He has a proven track record of building and scaling consulting practices and defining best practices across industries.



Core Competencies

- React
- Python
- Java
- Oracle
- Node.js
- Vue/PL/SQL
- Angular

Technical Skills & Platform Experience

- Databricks
- Cloudera
- BigQuery
- Streamsets
- Fivetran
- dbt
- AWS
- Tableau
- ServiceNow

Pete Peralta – Managing Director

Professional Summary

Pete is a seasoned digital leader with over 25 years of experience architecting comprehensive programs that drive business results. Prior to joining Cleartelligence, he held senior roles at Hero Digital, Accenture Interactive, and Publicis Sapient, where he led large-scale initiatives focused on Digital and AI transformation.

Pete brings deep expertise in defining transformation strategies, building cross-functional partnerships, and unlocking new pathways to enhance customer engagement.

His work in delivering complex digital engagements utilizing multiple technologies, platforms, and systems include brands like [Berkshire Hathaway Home Services](#), [Sprouts Farmers Market](#), [Southwest Airlines](#), and [The Rockefeller Foundation](#)



Core Competencies

- Executive Leadership
- Experience Design Strategy
- Digital Strategy & User Engagement
- Innovation & Emerging Technology
- Client Success

Keith Meyer – Practice Leader, Artificial Intelligence

Professional Summary

Keith Meyer is both a business and a technical strategist who has the rare ability to engage in business situations from the 50,000-foot level down to the 5-foot level. Keith has been the leader for many first-in-industry initiatives in tech, eCommerce and SaaS. At Cleartelligence, Keith works with clients to identify impactful data, applications and AI roadmaps. He follows through on recommendations by leading integrated teams in analytics, AI and software development projects.



Core Competencies

- Leadership
- Innovation & emerging Tech
- Defining and leading large and complex initiatives
- Systems architecture
- Data architecture
- Cloud architecture
- Software development

Technical Skills & Platform Experience

- Python, Java, C
- Advanced SQL
- Javascript, Typescript, React, Angular
- AI prompting & orchestration
- Data Integration tools & techniques
- AWS, GCP, Azure
- Docker, Kubernetes
- OpenStack
- Postgres, DB2, Oracle, SQL Server
- OpenAI, Gemini, Anthropic
- Snowflake
- Databricks
- Hadoop / Cassandra

Lisa Iyer – Principal Architect

Professional Summary

Lisa Iyer is a seasoned data leader with over 17 years of experience spanning data architecture, engineering, and business intelligence. She has a proven track record of designing and implementing scalable data frameworks that drive digital transformation initiatives across highly regulated industries including healthcare and financial services. Lisa brings deep expertise in modern cloud platforms including Databricks, Snowflake, and AWS, and is highly skilled in leading enterprise migration strategies to cloud-native solutions. A strong and collaborative leader, she excels at aligning data strategies with organizational goals to unlock measurable business value and innovation.



Core Competencies

- Cloud: Databricks, AWS (Redshift, Athena), Snowflake
- Data Engineering: DBT, Kafka, Talend, SQL Server
- BI & Analytics: Qlik Cloud, Qlik Sense, Tableau
- Languages: Python, SQL
- Data Catalog & Governance Tools: Atlan, Unity Catalog, Alation

Technical Skills & Platform Experience

- Qlik Sense Certified Data Architect
- Databricks Accredited Lakehouse Fundamentals
- Talend Core Certified Developer
- Talend Certified Solution Architect

Brian Bailey – Senior Consultant, Artificial Intelligence

Professional Summary

Brian Bailey is a senior consultant with deep expertise in AI/ML. With a Bachelor's and a Master's degree from MIT, Brian has a strong background in AI research and academia. He has also led client teams and architected successful GenAI solutions with proven ROI. Brian has experience building in cloud native environments, including AWS, Azure, and GCP. Beyond AI/ML, he has experience in general backend development, web services, database integration, data engineering, and data science.



Core Competencies

- GenAI/ML solution architecture
- Prototyping
- Client Relationship Management
- Innovation & emerging tech

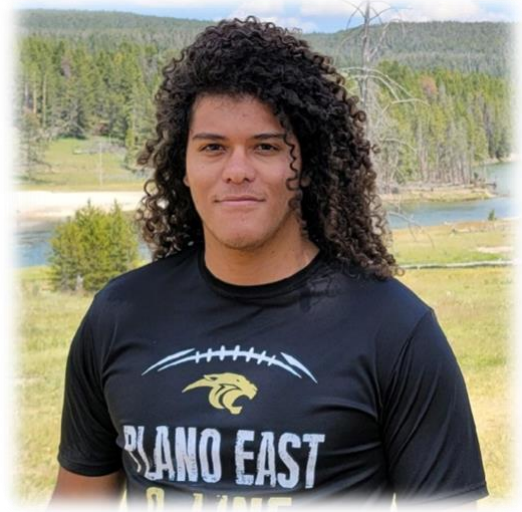
Technical Skills & Platform Experience

- Python
- LangChain
- PyTorch
- HuggingFace
- SQL
- OpenAI, Gemini, Claude

Mason Reiter – Consultant, Artificial Intelligence

Professional Summary

Mason Reiter is an AI consultant with over 3 years of professional experience with AI/ML. Having graduated from MIT and worked with Scale AI before his time at Cleartelligence, Mason brings a wealth of knowledge regarding the subject, especially regarding algorithm design and data management. He has experience creating real-world and synthetic training data for cutting edge LLMs such as Claude, Gemini, and Grok among others. At Cleartelligence, he architects AI strategies and solutions for clients. In addition to AI, he excels in cloud computing, data engineering, data science, and backend software development.



Core Competencies

- Generative AI & Machine Learning
- Data Strategy & Architecture
- Client Consulting & Solution Delivery
- Synthetic & Real-World Data Engineering
- Cloud-Native AI Development

Technical Skills & Platform Experience

- AI Prompt Engineering
- AI Agent Orchestration
- Python, Java, C++
- SQL
- PyTorch, HuggingFace
- Snowflake
- Databricks
- Gemini, OpenAI, Claude

Timeline and Budget

Engagement Timeline

The timeline has the engagement kicking off on December 2, 2025, and the AI Hub Data Commons Collaborative launching on September 28, 2026. A full PDF of the project plan is [available for download](#).



Figure 18 - Engagement Timeline

Budget Summary

The full breakdown of the budget by role is available in [Attachment C](#).

Services Budget

This budget utilized our U.S. based delivery team. All work will be done 100% onshore. We offer our onshore services as a blended rate of \$235 / hour.

Phase	Hours	Fee
Initiation & Design	6,220	\$1,461,700
Infrastructure & Setup	4,960	\$1,165,600
Development	10,197	\$2,396,295
Testing & Governance	2,639	\$620,165
Rollout & Adoption	1,650	\$387,750
Warranty	360	\$84,600
Total	26,026	\$6,116,110

Figure 19 - Budget by Project Phase

Budget Phase Justifications

Initiation & Design

This phase covers early-stage alignment activities, including project kickoff, stakeholder interviews, definition of architecture, and dataset curation criteria. The allocated hours and costs reflect the need to bring together diverse stakeholders, finalize governance models, and establish the foundation for the platform. These are critical investments to ensure that downstream technical and development work proceeds on solid footing and in compliance with MassTech's strategic and ethical AI goals.

Infrastructure & Setup

This budget line supports provisioning cloud infrastructure, identity and access management, and development/test environments. The effort and cost account for standing up secure environments that comply with state data protection requirements, ensuring that the DCC platform is scalable, reliable, and ready to host both curated and synthetic datasets. This is a one-time investment in the backbone of the Data Commons that underpins all subsequent development.

Development

Development is the most resource-intensive phase, covering platform buildout, dataset ingestion, metadata catalog development, synthetic data engine integration, and the creation of administrative dashboards. The hours and cost allocated here reflect the complexity of integrating multiple technologies (CKAN, synthetic data tools, Commons Credits ledger, API access) into a unified platform. This is the phase where most of the technical functionality is implemented and validated against the requirements.

Testing & Governance

This phase includes quality assurance (functional, performance, and security testing), as well as operationalizing the governance framework (RBAC, MFA, dataset tiering, audit logging, usage

agreements). The budget reflects the importance of ensuring trust, security, and fairness in the platform. These tasks require both technical validation and governance implementation to meet ethical AI and compliance standards mandated by MassTech.

Rollout & Adoption

The final phase covers user training, onboarding, documentation, and public launch. Costs here are tied to creating adoption materials, delivering stakeholder workshops, and ensuring that early users can successfully navigate the platform. This investment ensures that the Data Commons achieves its intended outcomes—driving adoption, collaboration, and AI innovation across Massachusetts.

Travel

In the Budget Template, we allocated \$17,500 quarter for a total engagement travel budget of \$70,000. This assumes 3-4 members of the team traveling to meeting with MassTech on a monthly basis. We look forward to refining this estimate to align to your preferred in person cadence.

NOTE: Travel (and any other expenses) are pass through costs and are not marked up in any way. Travel will align with MassTech's travel policies.

Software, Hosting, & SaaS Budget

To provide transparency into potential operating costs, the table below summarizes estimated annual expenses across three usage scenarios (Small, Medium, Large). These scenarios are based on key drivers such as dataset volume, user activity, and analytics intensity. Open-source platforms (e.g., CKAN, PortalJS, Superset) are included at no license cost, though they require hosting and support. Cloud infrastructure, Snowflake, identity management, translations, and analytics represent the primary cost variables.

- **Small Scenario** assumes a modest dataset footprint (<1 TB), limited concurrent users (<1,000), and basic translation/analytics needs.
- **Medium Scenario** reflects moderate adoption (10–25 TB, up to 10,000 users/month) with expanded translation, analytics, and observability.
- **Large Scenario** models significant growth (>100 TB, 50,000+ users/month) with enterprise-grade analytics and translation services.

Component	Small (est.)	Medium (est.)	Large (est.)
CKAN (core + required plugins)	Open Source	Open Source	Open Source
PortalJS (frontend)	Open Source	Open Source	Open Source
Strapi	\$5,000	\$5,000	\$5,000
AWS Hosting & Infrastructure (EC2, RDS, OpenSearch, S3, CloudFront, CloudWatch, Route 53)	\$12,000	\$40,000	\$120,000
Snowflake	\$10,000	\$35,000	\$100,000
Auth0	\$5,000	\$15,000	\$50,000
Metabase (SaaS Pro)	\$7,000	\$10,000	\$20,000
Bridge.xyz	See Licensing Notes	See Licensing Notes	See Licensing Notes
Translations.com GlobalLink (Project Director + Applanga)	\$4,000	\$6,000	\$6,000
Google Analytics (GA4 – Free Edition)	\$0	\$0	\$0
AI Fairness Tooling	Open Source	Open Source	Open Source
Total Annualized Estimate²	\$43,000	\$111,000	\$301,000

Figure 20 - License & Hosting Estimate

² The total cost outlay will also include the amount required to support the number of tokens to be issued for launch. MassTech will have the option of increasing / decreasing the amount in circulation and then will outlay or receive the funds that align with the change.

Security & Penetration Testing

We estimate the 3rd party penetration testing to range from \$15,000 - \$20,000 for the DCC. We can partner with a vendor of MassTech's choosing or recommend options for your consideration.

Licensing Notes

bridge.xyz

We will use bridge.xyz to issue a DCC stablecoin as the Commons Credit token. As stablecoin's require 1:1 currency backing the initial investment will be determined by the number of DCC tokens MassTech wants in circulation. MassTech will allocate the capital required for the coins that will be issued and bridge.xyz uses an Assets Under Management Fee of 30 basis points.

$$\text{Annualized bridge.xyz Fee} = \text{Issued Coins in Circulation} * 0.3\%$$

These costs can be offset since bridge.xyz pays interest to MassTech on the capital that it is holding. Currently, that interest is 4%.

Translations.com

Translations.com using multiple variables to determine pricing including number of languages, target languages, automated vs. human translation, and volume of words / sentences to translate. Once a word or phrase is translated and approved, it remains in the translation memory and future translations of that phrase do not have extra cost.

Our budget guidance assumes a medium size set of content (~10,000 words) that will be translated into three languages for the initial launch. We assumed Spanish, Chinese, and Arabic for the languages. Spanish was selected due to its common usage throughout the United States while Chinese and Arabic were chosen as the most expensive languages to translate. Reducing the number of languages, which languages are included, and/or the total volume of content will reduce the translation budget.

Additional Options to Manage Budget

We are providing the two alternative budget options below for MassTech's consideration to manage the overall budget for the engagement. Each option utilizes the same scope, approach, tasks, and effort. All lead roles are fulfilled onshore, so MassTech does not need to interact with resources in non-US time zones. If either of these options are selected, the same leadership as the fully onshore option will be in place for the steering committee, program, and technical leadership roles.

Offshore rates are offered at a blended rate of \$75 / hour. Both of the options provided below are included as tabs in [Attachment C](#).

Offshore QA

In this option, two of the QA execution roles will be staffed by our India based team. The QA lead role will be staffed in the US. **This option lowers the services delivery budget by \$192,000.**

Offshore QA + Limited Offshore Development

This option also staffs the same two QA roles from India. In addition, execution focused roles for application development, data engineering, and front-end development will also be staffed from our India team. The lead roles for each of these functions will be staffed by our US team. **This option lowers the services delivery budget by \$1,085,440.**

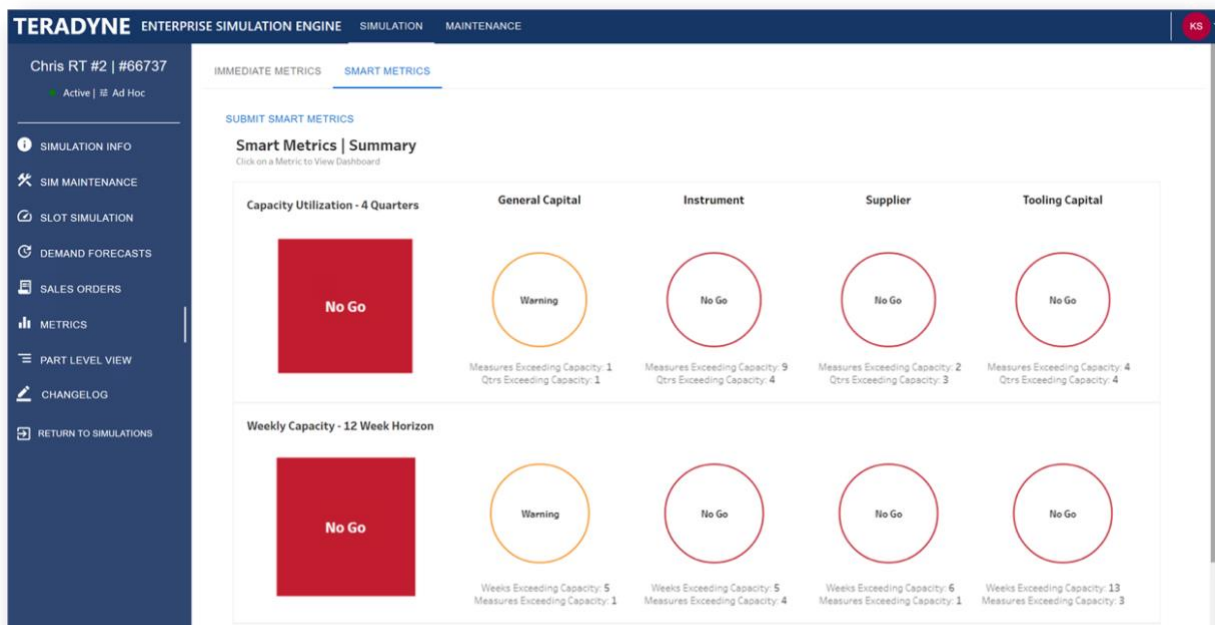
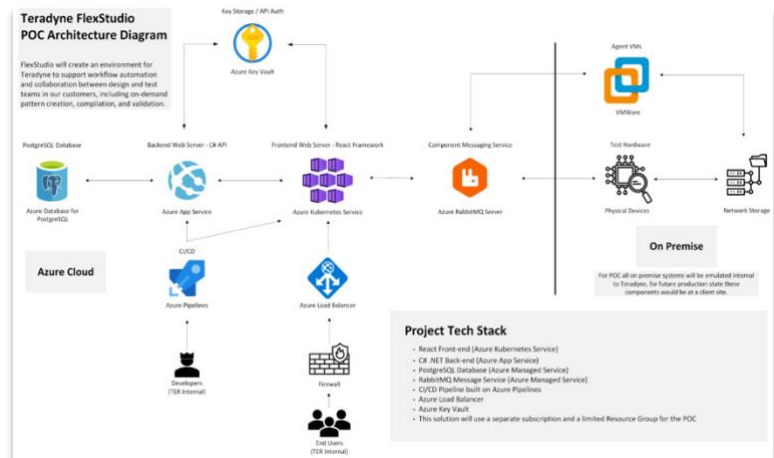
Case Studies

To highlight the breadth of our services and the alignment to the robust requirements of the DCC, we have selected the following use cases to highlight our experience working in complicated engagements with highly integrated systems.

Teradyne

[Teradyne](#) is a global supplier of automated test equipment (ATE) and industrial automation solutions, based in North Reading, MA. We developed a portal that allows users to upload specifications for what they need to build for a customer.

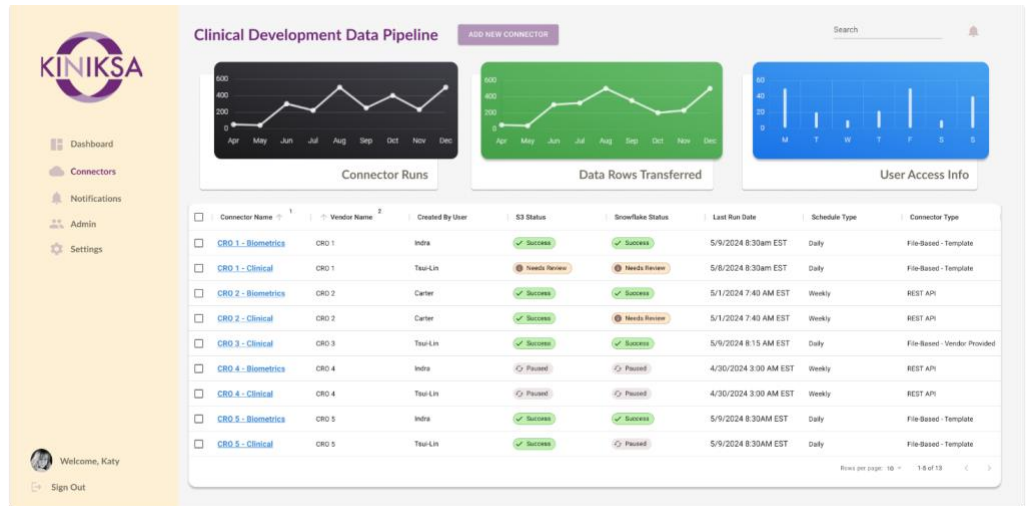
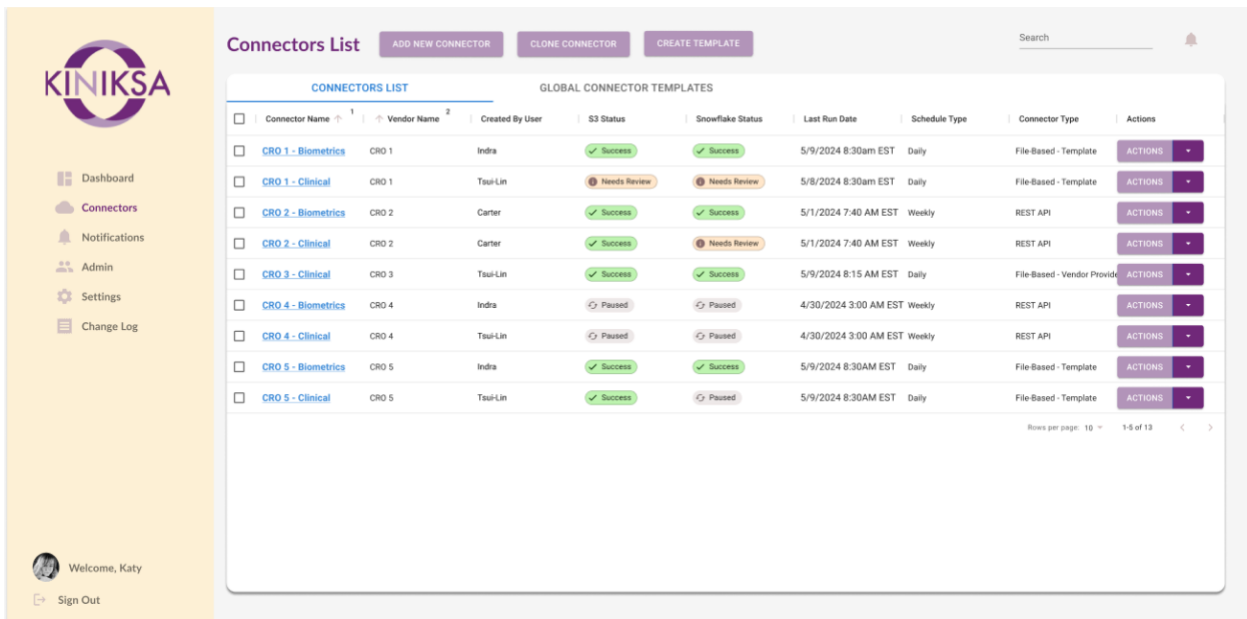
For each submission, the user can log in to a simulation experience to understand what time options would be for the build and subsequent delivery. If they don't find the timeline they are looking for, they can bump others out of a manufacturing slot and see what the impact would be. Impacts considered are not only work center time but also bill of materials and inventory/parts lead time. Once the implications of a change are understood, a committee gathers to review impacts and accept or reject individual changes.



Kiniksa

[Kiniksa Pharmaceuticals](#) is a biopharmaceutical company focused on discovering, developing, and commercializing therapies for patients with autoinflammatory and autoimmune diseases. They are based in Lexington, MA. We developed a configurable file intake and download system that enables the company to set-up partner companies (e.g. Contract Research Organizations) to upload clinical trials data. Partners are assigned API credentials and a pipeline codes for different inputs. When invoked by a partner, the system takes in the file, stores it, logs it, runs it through

validation, and invokes a configured pipeline to a target database destination. The partner company gets a success or error response. The API's allow for browsing history of file submissions. The system has internal management features like notification when an expected file is not received.

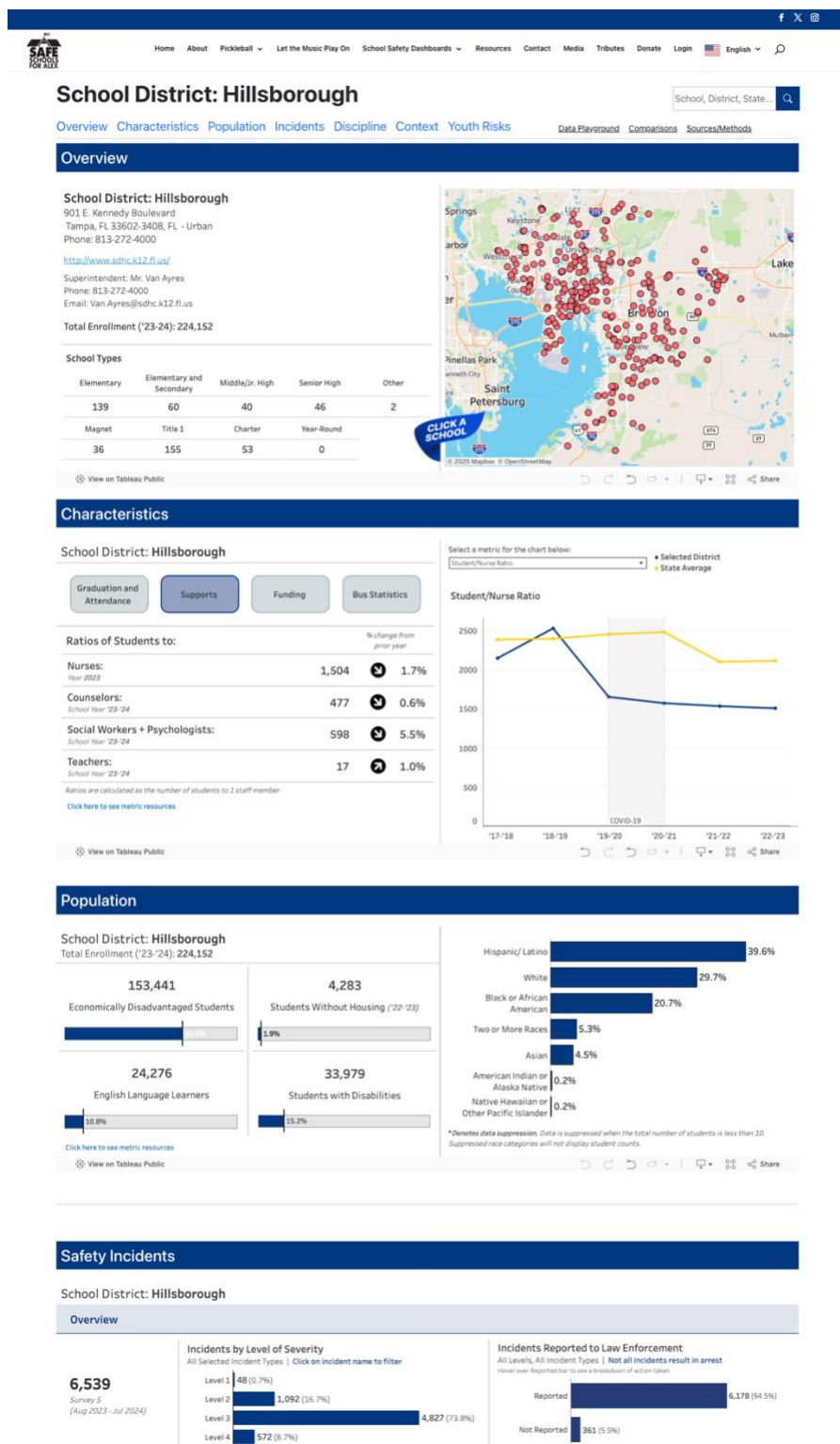
Connector Name	Vendor Name	Created By User	S3 Status	Snowflake Status	Last Run Date	Schedule Type	Connector Type	Actions
CRO 1 - Biometrics	CRO 1	Indra	Success	Success	5/9/2024 8:30am EST	Daily	File-Based - Template	ACTIONS
CRO 1 - Clinical	CRO 1	Tsui-Lin	Needs Review	Needs Review	5/8/2024 8:30am EST	Daily	File-Based - Template	ACTIONS
CRO 2 - Biometrics	CRO 2	Carter	Success	Success	5/1/2024 7:40 AM EST	Weekly	REST API	ACTIONS
CRO 2 - Clinical	CRO 2	Carter	Success	Needs Review	5/1/2024 7:40 AM EST	Weekly	REST API	ACTIONS
CRO 3 - Clinical	CRO 3	Tsui-Lin	Success	Success	5/9/2024 8:15 AM EST	Daily	File-Based - Vendor Provided	ACTIONS
CRO 4 - Biometrics	CRO 4	Indra	Paused	Paused	4/30/2024 3:00 AM EST	Weekly	REST API	ACTIONS
CRO 4 - Clinical	CRO 4	Tsui-Lin	Paused	Paused	4/30/2024 3:00 AM EST	Weekly	REST API	ACTIONS
CRO 5 - Biometrics	CRO 5	Indra	Success	Success	5/9/2024 8:30AM EST	Daily	File-Based - Template	ACTIONS
CRO 5 - Clinical	CRO 5	Tsui-Lin	Success	Paused	5/9/2024 8:30AM EST	Daily	File-Based - Template	ACTIONS

Safe Schools for Alex

[Safe Schools for Alex](#) is a nonprofit organization founded in memory of Alex Schachter, one of the 17 students and staff killed in the 2018 Marjory Stoneman Douglas High School shooting in Parkland, Florida. The organization is dedicated to improving school safety nationwide through data-driven research, advocacy, and resources for parents, educators, and policymakers. We have the honor of [supporting them in their mission](#) to ensure every child has a safe and supportive learning environment through their [School Safety Dashboard](#) initiative.

We are proud to play a part in the important mission of Safe Schools. Together we have created the following capabilities and impact:

- Establishing Uniform Data & Reporting for all school districts through the state of Florida
- School District Data available for all 50 states with continued expansion to cover all school districts in each state
- New Florida legislative school safety policies enabled, with Utah and other stated following
- Expanded engagement reach through Spanish translation of all web and dashboard experiences



Compliance and Acceptance

Attachment B

Massachusetts Technology Collaborative

Authorized Respondent's Signature and Acceptance Form

The undersigned is a duly authorized representative of the Respondent listed below. The Respondent has read and understands the RFP requirements. The Respondent acknowledges that all of the terms and conditions of the RFP are mandatory, and that Respondent's response is compliant with such requirements.

The Respondent understands that, if selected by the Mass Tech Collaborative, the Respondent and the Mass Tech Collaborative will execute an Agreement specifying the mutual requirements of participation.

Respondent agrees that the entire bid response will remain valid for sixty (60) days from receipt by the Mass Tech Collaborative.

I certify that Respondent is in compliance with all corporate filing requirements and State tax laws.

I certify that the statements made in this response to the RFP, including all attachments and exhibits, are true and correct to the best of my knowledge.

I acknowledge and understand the procedures for handling confidential materials submitted to the Mass Tech Collaborative as set forth in Section 3.1(d); I agree to be bound by those procedures and agree that the Mass Tech Collaborative shall not be liable under any circumstances for the disclosure of any materials submitted to the Mass Tech Collaborative pursuant to this RFP or upon the Respondent's selection.

Respondent: 
Kevin J. Lazorik

By: 
Kevin J. Lazorik

Name: Kevin J. Lazorik

Title: Chief Commercial Officer

Date: October 7, 2025

**Request for Taxpayer
Identification Number and Certification**

Go to www.irs.gov/FormW9 for instructions and the latest information.

Give form to the
requester. Do not
send to the IRS.

Before you begin. For guidance related to the purpose of Form W-9, see *Purpose of Form*, below.

Print or type. See Specific Instructions on page 3.	1 Name of entity/individual. An entry is required. (For a sole proprietor or disregarded entity, enter the owner's name on line 1, and enter the business/disregarded entity's name on line 2.) Clearintelligence Holdings, LLC	
	2 Business name/disregarded entity name, if different from above. Clearintelligence, LLC	
	3a Check the appropriate box for federal tax classification of the entity/individual whose name is entered on line 1. Check only one of the following seven boxes. ☐ Individual/sole proprietor ☐ C corporation ☐ S corporation ☐ Partnership ☐ Trust/estate ☒ LLC. Enter the tax classification (C = C corporation, S = S corporation, P = Partnership) P Note: Check the "LLC" box above and, in the entry space, enter the appropriate code (C, S, or P) for the tax classification of the LLC, unless it is a disregarded entity. A disregarded entity should instead check the appropriate box for the tax classification of its owner. ☐ Other (see instructions)	
	4 Exemptions (codes apply only to certain entities, not individuals; see instructions on page 3): Exempt payee code (if any) _____ Exemption from Foreign Account Tax Compliance Act (FATCA) reporting code (if any) _____ (Applies to accounts maintained outside the United States.)	
	3b If on line 3a you checked "Partnership" or "Trust/estate," or checked "LLC" and entered "P" as its tax classification, and you are providing this form to a partnership, trust, or estate in which you have an ownership interest, check this box if you have any foreign partners, owners, or beneficiaries. See instructions. ☐	
5 Address (number, street, and apt. or suite no.). See instructions. 275 Grove Street, #2-400		Requester's name and address (optional)
6 City, state, and ZIP code Auburndale, MA 02466		
7 List account number(s) here (optional)		

Part I Taxpayer Identification Number (TIN)

Enter your TIN in the appropriate box. The TIN provided must match the name given on line 1 to avoid backup withholding. For individuals, this is generally your social security number (SSN). However, for a resident alien, sole proprietor, or disregarded entity, see the instructions for Part I, later. For other entities, it is your employer identification number (EIN). If you do not have a number, see *How to get a TIN*, later.

Note: If the account is in more than one name, see the instructions for line 1. See also *What Name and Number To Give the Requester* for guidelines on whose number to enter.

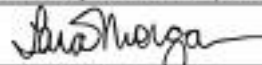
Social security number								
			-				-	
or								
Employer identification number								
9	2	-	1	0	6	5	6	3

Part II Certification

Under penalties of perjury, I certify that:

1. The number shown on this form is my correct taxpayer identification number (or I am waiting for a number to be issued to me); and
2. I am not subject to backup withholding because (a) I am exempt from backup withholding, or (b) I have not been notified by the Internal Revenue Service (IRS) that I am subject to backup withholding as a result of a failure to report all interest or dividends, or (c) the IRS has notified me that I am no longer subject to backup withholding; and
3. I am a U.S. citizen or other U.S. person (defined below); and
4. The FATCA code(s) entered on this form (if any) indicating that I am exempt from FATCA reporting is correct.

Certification instructions. You must cross out item 2 above if you have been notified by the IRS that you are currently subject to backup withholding because you have failed to report all interest and dividends on your tax return. For real estate transactions, item 2 does not apply. For mortgage interest paid, acquisition or abandonment of secured property, cancellation of debt, contributions to an individual retirement arrangement (IRA), and, generally, payments other than interest and dividends, you are not required to sign the certification, but you must provide your correct TIN. See the instructions for Part II, later.

Sign Here	Signature of U.S. person 	Date 02/14/2025
------------------	--	------------------------

General Instructions

Section references are to the Internal Revenue Code unless otherwise noted.

Future developments. For the latest information about developments related to Form W-9 and its instructions, such as legislation enacted after they were published, go to www.irs.gov/FormW9.

What's New

Line 3a has been modified to clarify how a disregarded entity completes this line. An LLC that is a disregarded entity should check the appropriate box for the tax classification of its owner. Otherwise, it should check the "LLC" box and enter its appropriate tax classification.

New line 3b has been added to this form. A flow-through entity is required to complete this line to indicate that it has direct or indirect foreign partners, owners, or beneficiaries when it provides the Form W-9 to another flow-through entity in which it has an ownership interest. This change is intended to provide a flow-through entity with information regarding the status of its indirect foreign partners, owners, or beneficiaries, so that it can satisfy any applicable reporting requirements. For example, a partnership that has any indirect foreign partners may be required to complete Schedules K-2 and K-3. See the Partnership Instructions for Schedules K-2 and K-3 (Form 1065).

Purpose of Form

An individual or entity (Form W-9 requester) who is required to file an information return with the IRS is giving you this form because they

Appendix

Index of Figures

Figure 1 - Program Delivery Structure	13
Figure 2 - Technical Architecture Overview	16
Figure 3 - CKAN Extensions Table.....	17
Figure 4 - Initial Role Mapping Table for RBAC.....	21
Figure 5 - Auth0 Integration Architecture	22
Figure 6 - Auth0 RBAC Mapping.....	23
Figure 7 - Credits Earned for Dataset Upload	27
Figure 8 - Spend Credits on Compute.....	28
Figure 9 - Snowflake Integration Architecture.....	31
Figure 10 - Dataset Upload with Snowflake Storage	33
Figure 11 - Log Archival & Retention	38
Figure 12 - Metabase Integration Architecture	40
Figure 13 - CKAN + Strapi - Combined Content Delivery.....	45
Figure 14 - Translation Reference Architecture	47
Figure 15 - AI Fairness Integration Architecture.....	51
Figure 16 - Data Management Platform Comparison	65
Figure 17 - Auth0 vs. Okta	67
Figure 18 - Engagement Timeline.....	82
Figure 19 - Budget by Project Phase	83
Figure 20 - License & Hosting Estimate.....	85

Links to Individual Files

- [Cleartelligence W9](#)
- [Bardess Group W9](#) (If needed)
- [Attachment C – Completed Budget Estimate Template](#)
- [Full Project Timeline \(PDF\)](#)
- [Staff Plan](#)

Key Assumptions & Dependencies

Solutions & Licensing

- Platform and License costs in the proposal are provided to indicate total cost of ownership of the solution. MassTech will procure and license the necessary solutions.

Experience & Branding

- MassTech will provide a branding guide (or similar) to ensure the visual design efforts align to the organizational brand standards.
- Accessibility and multilingual scope – The portal will meet WCAG 2.1 AA standards; initial locales (e.g., English plus one additional language) will be confirmed in the requirements phase.

Program Governance & Oversight

- MassTech will establish the Steering Committee and Advisory Group with clear decision SLAs to prevent sprint delays (e.g., 5 business days).
- Terms of Use, Privacy Policy, and dataset usage language will be provided and approved by Commonwealth legal counsel.
- Synthetic data governance – Subject Matter Experts (SMEs) will review fairness and privacy reports before synthesized datasets are published.
- Commons Credits policy – Earning/spending rules, quotas, and abuse protections will be defined with MassTech prior to launch.
- Training participation – MassTech will nominate administrators and end users for training, supporting a train-the-trainer model.
- Any scope or dataset additions beyond the confirmed initial list will be managed via the agreed change control process.
- MassTech will provide access to DCC compute or equivalent infrastructure for analytics and synthesis tasks.

Performance & Testing

- Availability and rate limits of external data sources are outside project control; features will adjust accordingly.
- Penetration testing and security reviews will include sufficient time for remediation before launch.
- Steering Committee sign-off during UAT will be required prior to production launch.

Industry Engagement Plan Framework

This framework serves as a starting point in our discussions with MassTech and the Steering Committee on engaging the key audiences for the DCC:

- AI Developers & Researchers
- Academic Institutions & Research Labs
- State government departments
- Civic tech groups
- AI Startups
- Enterprise AI Teams
- Non-profits working with Massachusetts open data

Goals

- Ensure a successful launch of the initial version of the DCC
- Grow Data Commons adoption across academic, industry, and government users.
- Build an ecosystem of data contributors and consumers in Massachusetts and beyond.

Engagement Activities

Requirements & Development

- Involve representative stakeholders across groups in the DCC requirements finalization to pressure test and validate defined approaches
- If any groups do not have proper representation in the steering committee, we can create an “Advisory Council” of ~10 people from across the different groups. This council would be included in formal demos and similar sessions to ask questions and provide feedback on the solution.

Pre-Launch

- The “Advisory Council” can be included in both the training and UAT processes
- Promotional messaging be defined and shared to support grassroots promotion across social channels

Post Launch

- Quarterly showcase events with industry and academic partners.
- Public launch campaign: press release, webinars, newsletter.
- Define Incentives such as free Commons Credits for early contributors. This can be done in several ways such as:
 - “Sponsored” Pilot program for 10 early adopters in Year 1.
 - “Double Credits” promotions that will increase the Commons Credits earned for key actions during a window
- Partnerships with incubator and accelerator programs that may include Commons Credits grants. Targets include:

- [Greentown Labs](#)
- [North Shore InnoVentures](#)
- [MassChallenge](#)
- [TechStars Boston](#)
- [The Engine](#)
- [Founder Institute](#)
- [Venture Development Center at UMass Boston](#)
- [Innovate @ BU](#)
- [LearnLaunch](#)
- [Roxbury Innovation Center](#)

Initial KPIs for Discussion

- 20+ datasets onboarded by end of Year 1.
- 1000+ registered users by end of Year 1.
- Net new registrations from across each target audience before the end of 2026

Sample Fairness Report

This sample fairness report demonstrates the approach the DCC will use to evaluate and document fairness considerations for datasets and models hosted on the platform.

Overview

Date: September 3, 2026

Prepared by: Massachusetts AI Hub – Data Commons Collaborative (DCC)

Dataset Under Review

Name: Example Health Outcomes Dataset

Domain: Healthcare

Size: 50,000 records, de-identified

Protected Attributes Considered: Gender, Race, Age Group

Methodology

- Dataset Auditing: Distribution analysis across protected groups.
- Bias Detection Metrics: Statistical parity difference, disparate impact ratio, equal opportunity difference.
- Model Fairness Evaluation: Performance of a logistic regression model predicting “Readmission Risk.”
- Bias Mitigation Techniques Applied: Reweighting (pre-processing), Equalized Odds post-processing.

Findings

Dataset Audit

Attribute	Group	% of Dataset	Notes
Gender	Male	52%	Balanced
Gender	Female	48%	Balanced
Race	White	70%	Over-represented
Race	Black	20%	Under-represented
Race	Other	10%	Small sample size
Age	<40	25%	Balanced
Age	40–65	50%	Balanced
Age	65+	25%	Balanced

Model Fairness Metrics (Pre-Mitigation)

Metric	Value	Interpretation
Statistical Parity Diff.	-0.18	Favoring majority group
Disparate Impact Ratio	0.72	Below 0.8 threshold (bias concern)
Equal Opportunity Diff.	-0.12	Minority group faces reduced opportunity

Mitigation Results

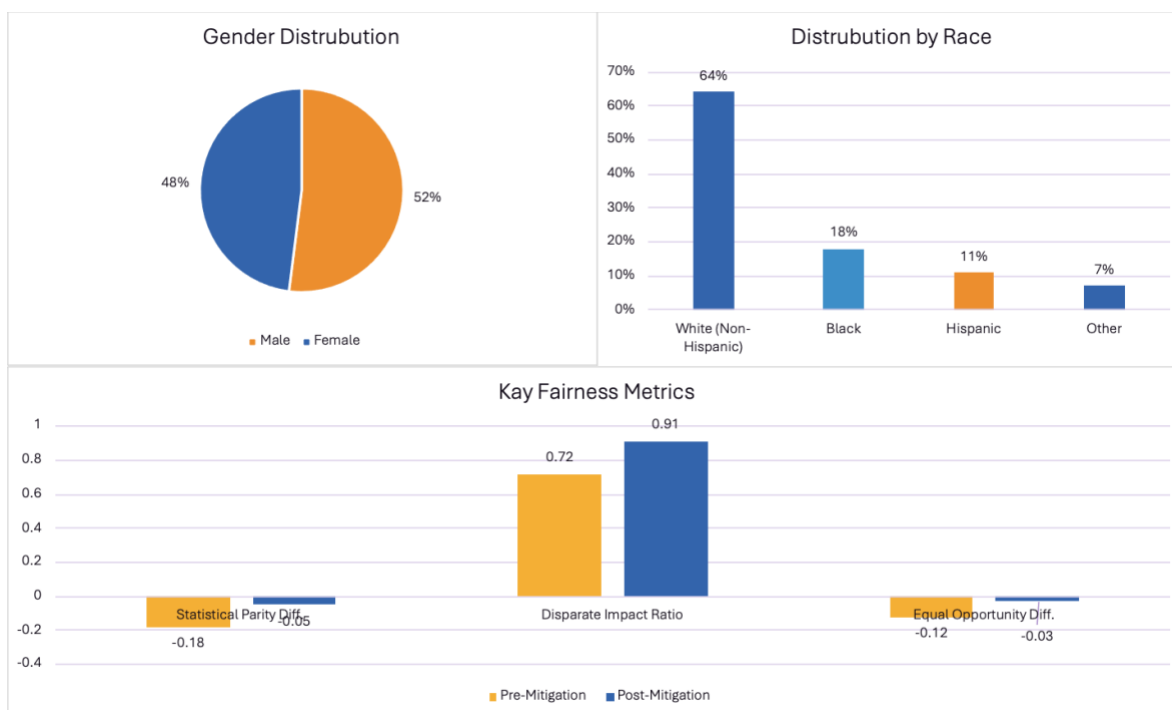
Metric (Post-Mitigation)	Value	Interpretation
Statistical Parity Diff.	-0.05	Improved fairness
Disparate Impact Ratio	0.91	Within acceptable range
Equal Opportunity Diff.	-0.03	Significant improvement

Fairness Impact Statement

- The dataset is broadly representative but has racial imbalance.
- Initial models showed measurable bias against under-represented groups.
- After applying bias mitigation techniques, fairness metrics improved across all measures, with minimal impact on overall model accuracy (-1%).

Recommendations

1. Continue to monitor dataset representation as new records are added.
2. Apply reweighing as part of the standard preprocessing pipeline.
3. Include this dataset in quarterly fairness drift monitoring.

Dashboard Screenshot (Illustrative)

Supporting Technical Information

Metadata Field Mapping

The following table outlines how CKAN's dataset metadata fields align with three widely recognized metadata standards: Dublin Core, Schema.org, and DCAT. This ensures that the Data Commons Collaborative (DCC) metadata management framework complies with FAIR data principles (Findable, Accessible, Interoperable, Reusable) and supports interoperability across academic, government, and industry data platforms.

CKAN Field	Dublin Core	Schema.org	DCAT
Dataset Title	dc:title	schema:name	dct:title
Description	dc:description	schema:description	dct:description
Author / Maintainer	dc:creator / dc:contributor	schema:creator / schema:contributor	dct:creator / dct:contributor
Organization (Publisher)	dc:publisher	schema:publisher	dct:publisher
Tags / Keywords	dc:subject	schema:keywords	dcat:keyword
License	dc:rights	schema:license	dct:license
Creation / Modification Date	dc:date	schema:dateCreated / schema:dateModified	dct:issued / dct:modified
Dataset Identifier (UUID/DOI)	dc:identifier	schema:identifier	dct:identifier
Dataset Type / Format	dc:format	schema:encodingFormat	dct:format / dcat:mediaType
Resource (Distribution)	dc:relation	schema:distribution → schema:contentUrl	dcat:Distribution / dcat:accessURL
Provenance / Source	dc:source	schema:isBasedOn	dct:source
Language	dc:language	schema:inLanguage	dct:language

Commons Credits Ledger Schema Example

Utilizing bridge.xyz natively addresses numerous requirements of the DCC including immutability, reporting, and loggings/auditability on the blockchain. Below is a sample JSON output of the transaction ledger for DCC Commons Credits, that are is natively available via the [bridge.xyz API](#).

Note that the public wallet address is stored as part of the user profile to align a user's issued credits with their DCC account.

```
{
  "transactions": [
    {
      "id": "txn_123456789",
      "chain": "solana",
      "stablecoin": {
        "symbol": "MASSDCC",
        "address": "Fz9yxx8Rv6Hhd42X6oGE4R9pQGLQgXkX9skYcbDqC7Vh"
      },
      "txHash": "5plu5y2JQG2HYzby1FkgNZbgnrkM89TZcY2i63zF4HT4ZvvhMxtQy9rD9S4T4A6r77ppJmYwTWr3zAqvhLWcLz8m",
      "from": "9oEpw6tCM6wdf5RaA4RoGpY3Fx2a2ec9hX6Y6oPFRUXS",
      "to": "7yb1EpGE9yz7s8eB7o9KgPbAf9eFhy1pJx7RtCR3aPzJ",
      "amount": "1000.00",
      "decimals": 6,
      "status": "confirmed",
      "type": "transfer",
      "createdAt": "2025-10-01T15:32:45Z",
      "confirmedAt": "2025-10-01T15:33:10Z",
      "metadata": {
        "memo": "Invoice #4321",
        "initiator": "api"
      }
    },
    {
      "id": "txn_987654321",
      "chain": "solana",
      "stablecoin": {
        "symbol": "MASSDCC",
        "address": "Fz9yxx8Rv6Hhd42X6oGE4R9pQGLQgXkX9skYcbDqC7Vh"
      },
      "txHash": "D6kM45QjK2MvrEwoKcYwSLvB4ux1GvQk78xrbhP27h9qzP4o3eNsudNjs2u8NsGq7r8fsEYmvXe1ZmukdUyXjY7",
      "from": "System",
      "to": "9oEpw6tCM6wdf5RaA4RoGpY3Fx2a2ec9hX6Y6oPFRUXS",
      "amount": "5000.00",
      "decimals": 6,
      "status": "confirmed",
      "type": "mint",
      "createdAt": "2025-10-02T10:21:13Z",
      "confirmedAt": "2025-10-02T10:21:59Z",
      "metadata": {
        "issuer": "Bridge.xyz",
        "notes": "Initial issuance"
      }
    }
  ],
  "pagination": {
    "page": 1,
    "pageSize": 25,
    "total": 2
  }
}
```

CMS + CKAN Content Model

This content model enables Strapi to manage editorial/narrative content while referencing authoritative datasets stored in CKAN. The design prioritizes: (a) separation of concerns, (b) minimal duplication of dataset metadata, (c) flexible editorial composition, (d) internationalization, (e) secure publishing workflows, and (f) future extensibility (e.g., additional headless CMS fields, portal widgets, and analytics).

The final Content Model will be determined during the [Detailed Requirements Gathering phase](#) of the engagement. The details provided below represent an expected starting state based on the RFP details.

Content Types (Overview)

The model uses a combination of Strapi Collection Types, Single Types, and Components.

- Collection Types
 - Pages
 - dataset_showcases
 - news_articles, guides
 - partners.
- Single Types
 - Home
 - site_settings.
- Core Components
 - rich_section
 - ckandataset_ref
 - media_gallery
 - call_to_action
 - external_link
 - person_card
 - seo

This structure allows editors to assemble rich pages while referencing CKAN datasets via stable CKAN IDs (package_name or UUID).

Collection Type Details

pages

Purpose: General purpose editorial pages composed of rich sections and optional dataset references.

Key Fields:

- title (string, required)
- slug (UID, unique, required) – based on title
- hero (component: rich_section)

- body (dynamic zone): [rich_section, media_gallery, call_to_action, ckandataset_ref]
- relatedDatasets (repeatable component: ckandataset_ref)
- seo (component: seo)
- locale (i18n)

Workflow: Draft & Publish enabled.

dataset_showcases

Purpose: Narrative pages that highlight a single CKAN dataset with contextual content, use cases, charts, and links.

Key Fields:

- title (string, required)
- slug (UID, unique, required)
- summary (text, required, 500 chars max)
- ckanDataset (component: ckandataset_ref – single, required)
- highlights (component: rich_section – repeatable)
- gallery (component: media_gallery – optional)
- relatedLinks (component: external_link – repeatable)
- seo (component: seo)
- locale (i18n)

Workflow: Draft & Publish enabled.

news_articles

Purpose: Announcements and updates (optionally referencing datasets).

Key Fields:

- title (string, required)
- slug (UID, unique, required)
- publishDate (date, required)
- author (component: person_card)
- body (dynamic zone): [rich_section, media_gallery, external_link, ckandataset_ref]
- tags (enumeration or relational taxonomies)
- seo (component: seo)
- locale (i18n)

guides

Purpose: How-to content (ingestion steps, governance policies, access requests).

Key Fields:

- title (string, required)
- slug (UID, unique, required)

- abstract (text, required)
- difficulty (enum: beginner|intermediate|advanced)
- steps (component: rich_section – repeatable)
- referencedDatasets (component: ckandataset_ref – repeatable)
- seo (component: seo)
- locale (i18n)

partners

Purpose: Directory of partner orgs contributing datasets, compute, or research.

Key Fields:

- name (string, required)
- slug (UID, unique, required)
- logo (media)
- website (string/url, validated)
- description (rich_section)
- contact (component: person_card – optional)
- featuredDatasets (component: ckandataset_ref – repeatable)
- seo (component: seo)

Massachusetts Technology Collaborative

**Grant Agreement
Standard Budget and Invoice Template
Invoiced Costs**

Line Number	Budgeting		
2	Applicant Information		
3	Applicant: Clearintelligence, LLC	Procurement No.:	
4	Budget Period:	Title of Proposed Project:	
5	Address:	MassTech Data Commons Collaborative	
6	MTC Funding Requested:		
7	Total Project Cost:		

LEGEND:

Formula cell

Data entry cell

Section completed for reference to showcase current planned effort

This section is for post-contract invoicing only. Do not complete with proposal.

[Full staff plan is also available for download](#)

Line Number	Cost Elements	Total Project	Invoiced Costs				Cumulative	
			Invoice 1	Invoice 2	Invoice 3	Invoice 4	Expended	Remaining
			Period: 12/2/2025 - 3/31/2026	Period: 4/1/2026 - 6/30/2026	Period: 7/1/2026 - 9/30/2026	Period: 10/1/2026 - 10/31/2026	Period: 12/2/2026 - 10/31/2026	
10	I. Direct Labor							
11	Track: Role (Location)	HRS Rate	HRS	HRS	HRS	HRS		
12	Program Management: Program Director (US)	1880 \$235 \$ 441,800.00	704 \$ 165,440	504 \$ 118,440.00	512 \$ 120,320.00	160 \$ 37,600.00	\$ 441,800.00	\$ -
13	Program Management: Sr. Business Analyst (US)	544 \$235 \$ 127,840.00	544 \$ 127,840	\$ -	\$ -	\$ -	\$ 127,840.00	\$ -
14	Program Management: Sr. Business Analyst (US)	1252 \$235 \$ 294,220.00	704 \$ 165,440	448 \$ 105,280.00	100 \$ 23,500.00	\$ -	\$ 294,220.00	\$ -
15	Program Management: Sr. Business Analyst (US)	1252 \$235 \$ 294,220.00	704 \$ 165,440	448 \$ 105,280.00	100 \$ 23,500.00	\$ -	\$ 294,220.00	\$ -
16	Program Management: Project Manager (US)	1660 \$235 \$ 390,100.00	704 \$ 165,440	504 \$ 118,440.00	452 \$ 106,220.00	\$ -	\$ 390,100.00	\$ -
17	Program Management: Project Manager (US)	1660 \$235 \$ 390,100.00	704 \$ 165,440	504 \$ 118,440.00	452 \$ 106,220.00	\$ -	\$ 390,100.00	\$ -
18	Website: Creative Director (US)	442 \$235 \$ 103,870.00	276 \$ 64,860	166 \$ 39,010.00	\$ -	\$ -	\$ 103,870.00	\$ -
19	Website: Sr. Visual Designer (US)	360 \$235 \$ 84,600.00	200 \$ 47,000	160 \$ 37,600.00	\$ -	\$ -	\$ 84,600.00	\$ -
20	Website: Sr. Visual Designer (US)	580 \$235 \$ 136,300.00	420 \$ 98,100	160 \$ 37,600.00	\$ -	\$ -	\$ 136,300.00	\$ -
21	Website: UX Lead (US)	300 \$235 \$ 70,500.00	180 \$ 42,300	120 \$ 28,200.00	\$ -	\$ -	\$ 70,500.00	\$ -
22	Website: Sr. Front End Developer (US)	612 \$235 \$ 143,820.00	220 \$ 51,700	392 \$ 92,120.00	\$ -	\$ -	\$ 143,820.00	\$ -
23	Website: Sr. Front End Developer (US)	272 \$235 \$ 63,920.00	\$ -	272 \$ 63,920.00	\$ -	\$ -	\$ 63,920.00	\$ -
24	External Data Sets: Sr. Application Development Engineer (US)	704 \$235 \$ 165,440.00	\$ -	384 \$ 90,240.00	320 \$ 75,200.00	\$ -	\$ 165,440.00	\$ -
25	External Data Sets: Senior Data Engineer (US)	384 \$235 \$ 90,240.00	\$ -	224 \$ 52,640.00	160 \$ 37,600.00	\$ -	\$ 90,240.00	\$ -
26	Hosted Data & Access: Sr. Application Development Engineer (US)	156 \$235 \$ 36,660.00	\$ -	156 \$ 36,660.00	0 \$ -	\$ -	\$ 36,660.00	\$ -
27	Hosted Data & Access: Lead Data Engineer (US)	214 \$235 \$ 50,290.00	214 \$ 50,290	\$ -	\$ -	\$ -	\$ 50,290.00	\$ -
28	Hosted Data & Access: Application Development Lead (US)	480 \$235 \$ 112,800.00	480 \$ 112,800	\$ -	\$ -	\$ -	\$ 112,800.00	\$ -
29	Hosted Data & Access: Sr. Application Development Engineer (US)	480 \$235 \$ 112,800.00	480 \$ 112,800	\$ -	\$ -	\$ -	\$ 112,800.00	\$ -
30	Hosted Data & Access: Sr. Application Development Engineer (US)	120 \$235 \$ 28,200.00	120 \$ 28,200	\$ -	\$ -	\$ -	\$ 28,200.00	\$ -
31	Commons Credits System: Application Development Lead (US)	472 \$235 \$ 110,920.00	372 \$ 87,420	100 \$ 23,500.00	\$ -	\$ -	\$ 110,920.00	\$ -
32	Commons Credits System: Application Development Engineer (US)	460 \$235 \$ 108,100.00	360 \$ 84,600	100 \$ 23,500.00	\$ -	\$ -	\$ 108,100.00	\$ -
33	Data Catalog & Metadata Mgmt: Lead Data Engineer (US)	1356 \$235 \$ 318,660.00	572 \$ 134,420	504 \$ 118,440.00	280 \$ 65,800.00	\$ -	\$ 318,660.00	\$ -
34	Data Catalog & Metadata Mgmt: Data Engineer (US)	1144 \$235 \$ 268,840.00	440 \$ 103,400	504 \$ 118,440.00	200 \$ 47,000.00	\$ -	\$ 268,840.00	\$ -
35	Dataset Curation: Lead Data Engineer (US)	712 \$235 \$ 167,420.00	674 \$ 158,390	36 \$ 8,460.00	\$ -	\$ -	\$ 167,420.00	\$ -
36	Dataset Curation: Senior Data Engineer (US)	856 \$235 \$ 201,160.00	520 \$ 122,200	336 \$ 78,960.00	\$ -	\$ -	\$ 201,160.00	\$ -
37	Dataset Curation: Senior Data Engineer (US)	800 \$235 \$ 188,000.00	520 \$ 122,200	280 \$ 65,800.00	\$ -	\$ -	\$ 188,000.00	\$ -
38	Synthetic Data Generation: Application Development Engineer (US)	504 \$235 \$ 118,440.00	\$ -	304 \$ 71,440.00	200 \$ 47,000.00	\$ -	\$ 118,440.00	\$ -
39	Synthetic Data Generation: Sr. Application Development Engineer (US)	1136 \$235 \$ 266,960.00	\$ -	464 \$ 109,040.00	512 \$ 120,320.00	160 \$ 37,600.00	\$ 266,960.00	\$ -
40	Synthetic Data Generation: Sr. Application Development Engineer (US)	464 \$235 \$ 109,040.00	\$ -	304 \$ 71,440.00	160 \$ 37,600.00	\$ -	\$ 109,040.00	\$ -
41	AI Fairness and Bias: Application Development Engineer (US)	544 \$235 \$ 127,840.00	\$ -	464 \$ 109,040.00	80 \$ 18,800.00	\$ -	\$ 127,840.00	\$ -
42	AI Fairness and Bias: Application Development Engineer (US)	344 \$235 \$ 80,840.00	\$ -	264 \$ 62,040.00	80 \$ 18,800.00	\$ -	\$ 80,840.00	\$ -
43	AI Fairness and Bias: Data Engineer (US)	424 \$235 \$ 99,640.00	\$ -	344 \$ 80,840.00	80 \$ 18,800.00	\$ -	\$ 99,640.00	\$ -
44	Security & Access Control: Sr. Application Development Engineer (US)	664 \$235 \$ 156,040.00	\$ -	464 \$ 109,040.00	200 \$ 47,000.00	\$ -	\$ 156,040.00	\$ -
45	Security & Access Control: Application Development Lead (US)	764 \$235 \$ 179,540.00	80 \$ 18,800	484 \$ 113,740.00	200 \$ 47,000.00	\$ -	\$ 179,540.00	\$ -
46	Quality Assurance: Quality Assurance Lead (US)	830 \$235 \$ 196,950.00	\$ -	384 \$ 90,240.00	406 \$ 95,410.00	40 \$ 9,400.00	\$ 196,950.00	\$ -
47	Quality Assurance: Quality Assurance Test Engineer (US)	464 \$235 \$ 109,040.00	\$ -	344 \$ 80,840.00	120 \$ 28,200.00	\$ -	\$ 109,040.00	\$ -
48	Quality Assurance: Quality Assurance Test Engineer (US)	736 \$235 \$ 172,960.00	\$ -	344 \$ 80,840.00	392 \$ 92,120.00	\$ -	\$ 172,960.00	\$ -
49	Total Direct Labor	\$ 6,116,110.00	10,192 \$ 2,395,120.00	10,528 \$ 2,491,580.00	4,946 \$ 1,179,810.00	360 \$ 84,600.00	\$ 6,116,110.00	\$ -
50	I. Direct Labor Fringe Cost	Rate (%):	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
51	II. Subcontractors/Consultants/Contracted Staffing							
52	Firm name/type of services to be provided	HRS Rate	HRS	HRS	HRS	HRS		
53	N/A		\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
54			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
55			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
56			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
57			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
58			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
59			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
60			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
61			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
62			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
63			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
64	Total Subcontractors/Consultants		\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
65	III. Direct Materials (provide description and costs basis)							
66	Curriculum Development & Training		\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
67			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
68			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
69			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
70			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
71			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
72			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
73			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
74	Total Direct Materials		\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
75	Travel Budget Assuming a roughly 1 onsite visit / month for steering committee and Clearintelligence leads	\$70,000	\$17,500	\$17,500	\$17,500	\$17,500	\$70,000	\$0
76	V. Other Direct Costs (list type and description such as venue rental costs and cost basis)							
77			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
78			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
79			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
80			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
81			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
82			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
83			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
84	Total Other Direct Costs		\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
85	VI. General & Administrative Expense/Overhead	Rate (%):	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
86			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
87			\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
88	Total PROJECT COSTS		\$6,186,110	\$2,412,620	\$2,491,580	\$1,179,810	\$102,100	\$6,186,110